LittleMissFits: Ein Game-With-A-Purpose zur Evaluierung subjektiver Verständlichkeit von latenten Faktoren in Empfehlungssystemen

Johannes Kunkel, Benedikt Loepp, Esther Dolff, Jürgen Ziegler
University of Duisburg-Essen, Duisburg, Germany
{firstname.lastname}@uni-due.de

ZUSAMMENFASSUNG

Empfehlungssysteme, die mit Hilfe latenter Faktormodelle Empfehlungen generieren, arbeiten äußerst genau und sind entsprechend weit verbreitet. Da die Berechnung der Empfehlungen jedoch auf der statistischen Auswertung von Benutzerbewertungen basiert, gestaltet es sich schwierig, die Empfehlungen dem Nutzer gegenüber zu erklären. Daher werden die Systeme häufig als intransparent wahrgenommen und können oft ihr volles Potential nicht entfalten. Erste Ansätze zeigen allerdings, dass die latenten Faktoren solcher Modelle semantische Eigenschaften der Produkte widerspiegeln. Dabei ist bislang unklar, ob die zum Teil sehr komplexe Parametrisierung, die z.B. die Anzahl der Faktoren festlegt, Auswirkungen auf die semantische Verständlichkeit hat. Da dies sehr von der subjektiven Wahrnehmung abhängt, präsentieren wir mit *LittleMissFits* ein Online-Spiel, das es erlaubt, mittels Crowd-Sourcing die Konsistenz der latenten Faktoren zu untersuchen. Die Ergebnisse einer Nutzerstudie mit diesem Spiel zeigen, dass eine höhere Anzahl von Faktoren das Modell weniger verständlich erscheinen lässt. Darüber hinaus fanden sich Unterschiede innerhalb der Faktormodelle bezüglich der Verständlichkeit der einzelnen Faktoren. Zusammengenommen stellen die Ergebnisse eine wertvolle Grundlage dar, um künftig die Transparenz entsprechender Empfehlungssysteme zu steigern.

KEYWORDS

Empfehlungssysteme, Matrixfaktorisierung, Game-with-a-Purpose, Transparenz, Nutzerkontrolle

ACM Reference Format:
Johannes Kunkel, Benedikt Loepp, Esther Dolff, Jürgen Ziegler. 2019. LittleMissFits: Ein Game-With-A-Purpose zur Evaluierung subjektiver Verständlichkeit von latenten Faktoren in Empfehlungssystemen. In *Proceedings of the Mensch und Computer 2019 Workshop Gam-R – Gamification Reloaded.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.18420/muc2019-ws-576

MuC'19 Workshops, Hamburg, Deutschland

https://doi.org/10.18420/muc2019-ws-576

1 EINLEITUNG

Empfehlungssysteme (ES) werden eingesetzt, um Nutzern proaktiv personalisierte Produktvorschläge zu präsentieren. Die dazu verwendeten Algorithmen sind über die Jahre immer genauer in der Vorhersage der Interessen der Nutzer geworden. Allerdings geht eine hohe Genauigkeit nicht zwangsläufig mit gesteigerter Zufriedenheit der Nutzer einher [7]. Vielmehr sind nutzerzentrierte Aspekte wie die Verständlichkeit der Empfehlungen ebenfalls von großer Bedeutung. So werden bei modernen ES häufig textuelle Erklärungen zu den Empfehlungen präsentiert. Dies gestaltet sich jedoch mitunter als schwierig, wenn abstrakte Modelle zur Empfehlungsgenerierung zum Einsatz kommen. Solche modellbasierten Verfahren wie z.B. die Matrixfaktorisierung (MF) [8] sind zwar für ihre hohe objektive Empfehlungsqualität bekannt, eignen sich aufgrund ihrer intransparenten algorithmischen Arbeitsweise jedoch kaum, um die Empfehlungen für den Nutzer verständlich zu erklären.

Zuletzt hat die Forschung im Bereich um transparentere ES zugenommen [22]. So wird versucht, Erklärungen auch in solchen Ansätzen bereitzustellen, die Empfehlungen mit Hilfe von MF gelernten latenten Faktormodellen generieren. Obwohl die Faktoren solcher Modelle aufgrund ihrer rein statistischen Berechnung zunächst keinen realweltlichen Eigenschaften zugeordnet sind, konnte eine nachträgliche Zuordnung von Faktoren zu semantische Dimensionen erfolgreich hergestellt werden [12, 13, 18]. Diese semantischen Dimensionen können z.B. mit Tags, die andere Nutzer bezüglich der zugrundeliegenden Produkte vergeben haben, in Verbindung gebracht werden, um Empfehlungen zu erklären [13]. Die Grundannahme solcher Ansätze ist dabei nicht nur, dass semantische Dimensionen in den Faktormodellen grundlegend existieren, sondern auch, dass diese innerhalb der Faktoren konsistent bzw. zwischen den Faktoren unterschiedlich sind. Eine hohe Trennschärfe zwischen den Faktoren ist z.B. eine wichtige Voraussetzung, um eindeutig

erklären zu können, aufgrund welcher Kriterien eine Empfehlung generiert wurde.

Bislang wurde jedoch weder die empfundene Konsistenz innerhalb der latenten Faktoren noch die wahrgenommene Trennschärfe zwischen den Faktoren explizit untersucht. Auch wurde noch kein Augenmerk darauf gelegt, welchen Einfluss die Parametrisierung der MF-Algorithmen auf die resultierenden Dimensionen hat. Ein Parameter, der Auswirkungen auf die von einem Faktormodell transportierte Semantik haben könnte, ist z.B. die Anzahl der latenten Faktoren mit der das Modell trainiert wird.

Zusammengefasst sind wir daher an der Beantwortung folgender Forschungsfragen interessiert:

- *FF1* Werden Faktoren der MF inhaltlich als semantisch konsistent wahrgenommen?
- *FF2* Werden Faktoren untereinander als semantisch unterschiedlich wahrgenommen?
- *FF3* Hat die Parametrisierung der Faktormodelle — insbesondere die Anzahl der Faktoren — einen Einfluss auf deren wahrgenommene Verständlichkeit?

Um diese Forschungsfragen zu untersuchen, präsentieren wir in diesem Beitrag das Game-with-a-Purpose [24] *LittleMissFits*. Bei diesem Online-Spiel wählt der Spieler aus einer Auswahl von Filmen, die mit Hilfe von Filmplakat und Metadaten präsentiert werden, denjenigen aus, der sich semantisch von den übrigen Filmen zu unterscheiden scheint. Der zu findende Film ist dabei bezüglich eines bestimmten latenten Faktors besonders stark ausgeprägt, die übrigen dargestellten Filme bezüglich eines anderen. Wird das Spiel häufig genug gespielt, lässt sich anhand der Anzahl erfolgreich gespielter Runden ablesen, wie die Konsistenz innerhalb der Faktoren und die Verschiedenheit zwischen diesen wahrgenommen werden. Eine erste Nutzerstudie ($N = 46$) liefert positive Resultate bezüglich des empfundenen Spielspaßes und weist auf eindeutige Unterschiede bezüglich der genannten Aspekte hin. Generell zeigt sich zudem, dass eine geringere Faktoranzahl zu besserer Verständlichkeit führt. Die hier vorgestellten Ergebnisse können dabei helfen, Faktormodelle der MF künftig verständlicher darzustellen, Erklärungen zu deren Empfehlungen zu generieren und somit für eine höhere Transparenz in modellbasierten ES zu sorgen. Zudem zeigen die Ergebnisse beispielhaft, wie ein an sich sehr abstraktes, algorithmisches Problem veranschaulicht und spielerisch gelöst werden kann.

2 VERWANDTE ARBEITEN

Die Generierung von Empfehlungen findet heutzutage meistens vollautomatisch auf Basis von Bewertungen statt, welche die Nutzer für die Produkte abgegeben haben. Populär ist insbesondere die als *Collaborative Filtering* (CF) bekannte Methode, welche beispielsweise bei Amazon [20] und Netflix [3] eingesetzt wird. Moderne CF-Algorithmen arbeiten selten direkt auf den Produktbewertungen, sondern lernen aus diesen Daten zunächst ein abstraktes Modell, welches anschließend für die Generierung von Empfehlungen verwendet wird. Ein prominentes Beispiel für eine solche modellbasierte CF-Technik ist die *Matrixfaktorisierung* (MF) [8]: Aus der Matrix aller abgegebenen Bewertungen der Nutzern für die Produkte werden mit Hilfe eines Optimierungsverfahrens latente Faktoren abgeleitet. Resultat ist ein Faktormodell, das für jeden Nutzer und jedes Produkt einen Vektor beinhaltet, der ausdrückt, wie sehr ein Nutzer an den Faktoren interessiert ist bzw. wie stark diese in einem Produkt ausgeprägt sind. Je nach Anwendungsfall werden 2 bis über 100 solcher Faktoren berechnet. Wird der Faktor-Vektor eines Nutzers mit dem eines Produkts multipliziert, ergibt sich eine Vorhersage dafür, wie dieser Nutzer dieses Produkt bewerten würde. MF ist dafür bekannt, dass diese Vorhersagen sehr akkurat sind, weshalb das Verfahren überaus verbreitet ist.

Neben der Fähigkeit, möglichst genaue Empfehlungen zu berechnen, werden moderne ES immer häufiger daran gemessen, wie gut sie andere Nutzerbedürfnisse erfüllen [6, 7]. So hat sich gezeigt, dass Nutzer sich wünschen die Empfehlungen und die Gründe dahinter besser zu verstehen [22]. Diese Eigenschaft wird als Transparenz eines ES bezeichnet und kann beispielsweise durch textuelle Erklärungen zu den Empfehlungen gesteigert werden. Eine einfache Form der textuellen Erklärung von CF-basierten Empfehlungen findet sich bei Amazon: „Nutzer, die … kauften, kauften auch …". Wenn modellbasierte Techniken wie z.B. MF zum Einsatz kommen, können Erklärungen jedoch nur schwerlich generiert werden. Dies liegt insbesondere daran, dass die latenten Modelle statistisch erlernt werden und die semantische Bedeutung der Faktoren a priori unbekannt ist.

Werden semantische Dimensionen hinter den latenten Faktoren aufgedeckt, ist jedoch auch in modellbasierten ES Potenzial vorhanden, aufschlussreiche Erklärungen für die Empfehlungen bereitzustellen. In verschiedenen Arbeiten wurde bereits gezeigt, dass sich die Faktoren erfolgreich mit inhaltlichen Daten verbinden lassen. Hierbei kamen beispielsweise angepasste MF-Varianten zum Einsatz um eine Verbindung von nutzervergebenen Tags zu den Faktoren herzustellen [13]. In anderen Ansätzen wurden die semantischen Dimensionen hinter den latenten Faktoren nach dem Erlernen des Modells aufgedeckt. Beispielsweise können die Ausprägungen der Faktoren für einzelne Produkte visualisiert [16] oder gesamte Produkträume auf Basis der latenten Faktoren dargestellt werden [2, 10]. Ebenso positive Erfahrungen konnten mit crowd-basierten Ansätzen gemacht werden, bei denen Nutzer die Bedeutung hinter den latenten Faktoren identifizieren [11, 12]. Der Einsatz von *Games-with-a-Purpose* (GWAP) scheint hierfür besonders geeignet zu sein.

In GWAP erledigen Spieler Aufgaben, die nicht oder nur unzureichend von einem Computer durchgeführt werden

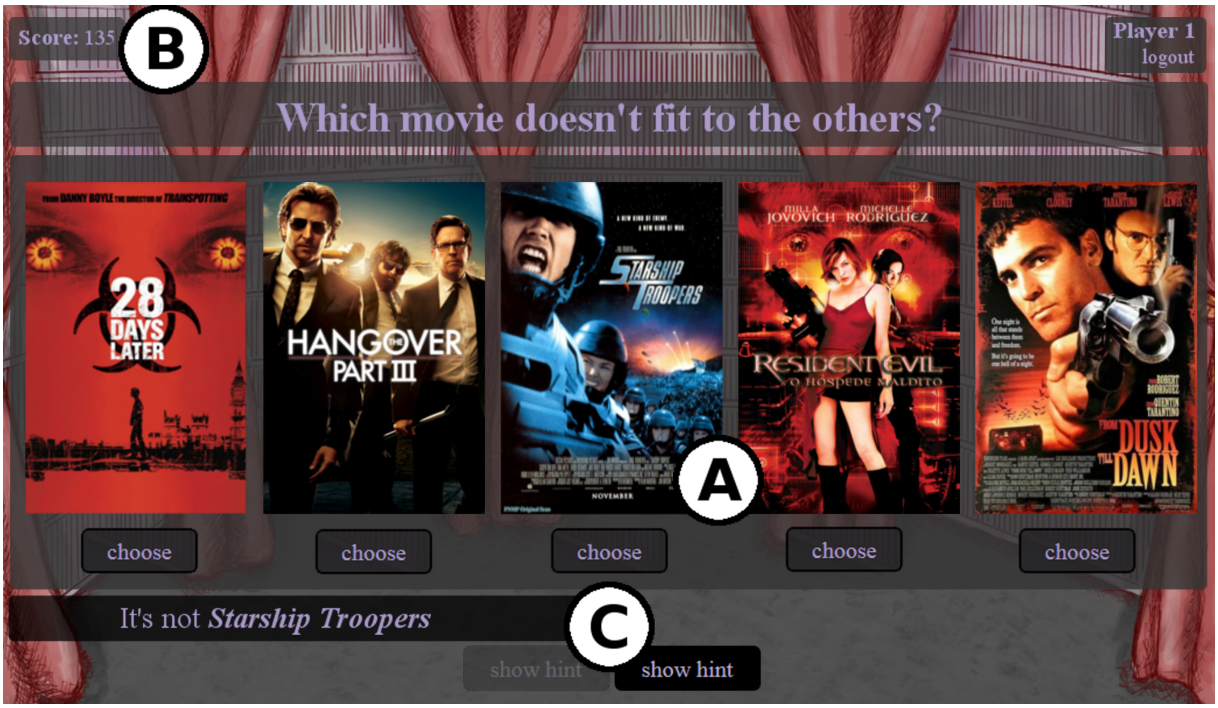

Abbildung 1: Hauptbildschirm von *LittleMissFits*. Aus einer Auswahl von fünf Filmen (A) muss derjenige identifiziert werden, der nicht zu den anderen passt. Für jeden korrekt identifizierten Film, wird der Punktestand des Spielers erhöht (B). Zwei zusätzliche Hinweise stehen zur Verfügung, die jeweils einen der Filme von der Auswahl ausschließen (C). Für jeden verwendeten Hinweis wird der Punktestand des Spielers reduziert.

können. Eines der prominentesten Beispiele ist das durch von Ahn [23] entwickelte *ESP Game*: Zwei Spieler überlegen, welchen Begriff der jeweilige Mitspieler für die Beschreibung eines gegebenen Bildes eingeben würde. Das Ziel einer Spielrunde ist erreicht, wenn beide Spieler unabhängig voneinander denselben Begriff eingegeben haben. Entworfen wurde das Spiel, welches später unter dem Namen „Google Image Labeler" weiter betrieben wurde, zu dem Zweck, nicht näher beschriebene Bilder für Suchmaschinen zugänglich zu machen. Gleiches Ziel verfolgt auch *KissKissBan* von Ho et al. [5]. Hier fungiert neben den zwei Spielern, die versuchen gemeinsame Stichwörter zu kreieren, eine dritte Person als Gegenspieler, der zu Spielbeginn so viele unerlaubte Wörter wie möglich eingibt. Durch diese Spielmechanik wurden bei *KissKissBan* diversere Begriffe eingegeben als beim *ESP Game*. Neben der Beschriftung von Bildern sind GWAP auch schon für andere Aufgaben genutzt worden, beispielsweise um textuelle Mehrdeutigkeiten aufzulösen [19], aber auch im Bereich von ES, etwa zur Erhebung von Präferenzen [1, 4, 21] oder zur Bestimmung von Ähnlichkeiten zwischen Produkten [25].

Wie zuvor erwähnt, wurden GWAP ebenfalls bereits erfolgreich für das Finden von beschreibenden Begriffen für die latenten Faktoren eines mit Hilfe von MF gelernten Modells angewandt [11, 12]. Hierbei bleibt jedoch die Frage unbeantwortet, welchen Einfluss die Parametrisierung, mit der das Faktormodell gelernt wurde, auf die Verständlichkeit der latenten Faktoren hat. Ebenso bleibt unklar, wie sehr die Faktoren innerhalb als konsistent und untereinander als unterschiedlich wahrgenommen werden — obwohl dies entscheidende Aspekte sind, wenn Erklärungen basierend auf den latenten Modellen generiert werden sollen. In diesem Beitrag präsentieren wir ein GWAP, um diesen Fragen zu begegnen.

3 LITTLEMISSFITS

Das hier vorgestellte GWAP *LittleMissFits*[1] (Abbildung 1), stellt einen crowd-basierten Ansatz dar, um Eigenschaften wie wahrgenommene Zusammengehörigkeit und Diversität in latenten Faktormodellen der MF zu erheben. Die Methode ist dabei in zwei Phasen gegliedert: Eine Offline-Phase und eine Online-Phase. In der Offline-Phase, wird zunächst ein latentes Faktormodell auf Basis von Filmbewertungen berechnet. Anschließend werden repräsentative Filme für jeden Faktor ermittelt. In der Online-Phase wird das Spiel durch

[1] http://lmf.ci.interactivesystems.info

Abbildung 2: Zu Beginn des Spiels werden Spieler in die narrative Hintergrundgeschichte zu *LittleMissFits* eingeführt.

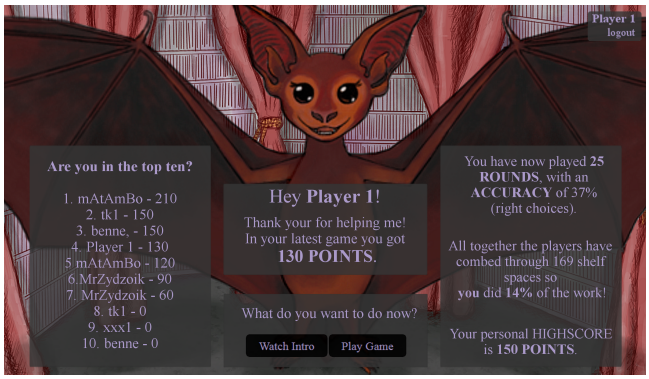

Abbildung 3: Statistikbildschirm von *LittleMissFits*. Hier befinden sich motivierende Elemente, wie einer Rangliste mit Highscores, einer Anzeige zur persönlichen Bestleistung und dem eigenen Anteil am Gesamtfortschritt.

Nutzer gespielt. *LittleMissFits* ist als Webseite umgesetzt, um möglichst vielen Spielern den Zugriff zu ermöglichen. Anhand der Spieldaten, können neben der wahrnehmbaren inhaltliche Einheitlichkeit der Faktoren auch Faktormodelle unterschiedlicher Parametrisierung evaluiert und Auswirkungen der Parameter auf die Verständlichkeit der latenten Faktoren untersucht werden.

Spielprinzip

Zentrale Herausforderung für Spieler von *LittleMissFits* besteht in der erfolgreichen Identifikation eines unpassenden Films aus fünf Alternativen (Abbildung 1). Im Hintergrund besteht diese Gruppe aus Filmen, die zu zwei zufällig bestimmten Faktoren gehören. Vier der fünf Filme stammen dabei von einem Faktor des Faktormodells (nachfolgend als *Faktor A* bezeichnet), während der fünfte Film von einem anderen Faktor stammt (nachfolgend als *Faktor B* bezeichnet). Die fünf Filme werden dem Spieler angezeigt, ohne die Informationen über die Zugehörigkeit zu den Faktoren. Als Hilfe können jedoch Informationen wie Poster, Liste mit Schauspielern und Regisseuren, Inhaltsbeschreibung und Trailer angeschaut werden. Der Spieler soll nun den Film finden, der seiner Ansicht nach nicht zu den anderen passt. Anhand der Quote mit der Spieler den Film von *Faktor B* als unpassend markieren, können Aussagen über die eingesetzten latenten Faktoren gemacht werden. Wenn das Spiel z.B. für bestimmte Faktoren zu hohen Erfolgsquoten führt, deutet dies auf eine wahrnehmbare Einheitlichkeit der Filme aus *Faktor A* aber auf eine gute Unterscheidbarkeit zu Filmen aus *Faktor B* hin.

Spielelemente

Da in der Vergangenheit gezeigt wurde, dass narrative Elemente einen großen Effekt auf die Motivation von Spielern an GWAP teilzunehmen haben können [17], wurde eine Hintergrundgeschichte in *LittleMissFits* implementiert. Diese besteht in dem fiktiven in-game Charakter *Liz*, welche erzählt ihre Filmsammlung nach semantischen Inhalten und nicht alphabetisch zu sortieren (Abbildung 2). Da diese Sortierung häufig durcheinander gerät, sollen die Spieler nun helfen die Ordnung wiederherzustellen, indem inhaltlich unpassende Filme markiert werden, was gleichzeitig der Hauptaufgabe von *LittleMissFits* entspricht.

Neben narrativen Elementen, setzt *LittleMissFits* auf typische Spielelemente, die sich in bisherigen GWAP bewährt haben [11, 24]. Beispiele hierfür sind Punkte und kompetitive Bestenlisten. Punkte können Spieler in *LittleMissFits* durch korrekt identifizierte Filme erlangen (Abbildung 1, B). Wird der korrekte Film gefunden erhält der Spieler 100 Punkte. Wird ein falscher Film ausgewählt, werden dem Punktestand 50 Punkte abgezogen. Um den Nutzern die Auswahl zu erleichtern, können zwei Hinweisfelder einzeln aufgedeckt werden. Diese schließen jeweils einen Film als Ausreißer aus (Abbildung 1, C). Die Verwendung dieser Hinweise, muss jeweils mit einem Abzug von 15 Punkten bezahlt werden. Ein Spiel in *LittleMissFits* setzt sich aus 5 Runden zusammen, für die kein Zeitlimit besteht. Anschließend wird eine Seite mit Statistiken angezeigt (Abbildung 3), die neben einer Highscore-Liste und der persönlichen Bestleistung auch den eigenen Anteil an den insgesamt gespielten Runden beinhaltet. Hiermit werden kollaborative Motivationsaspekte adressiert, um neben solchen Spielern, die den Wettkampf suchen (vgl. *Conqueror* [15]), auch Spieler zu motivieren, die Motivation besonders durch soziale Spielelemente empfinden (vgl. *Socialiser* [15]).

Datengrundlage

LittleMissFits basiert auf dem *20M-Datensatz* von *Grouplens*[2], welcher 20 Millionen Bewertungen von 137 000 Nutzern für 27 000 Filme enthält. Informationen zu den Filmen wurden

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

Tabelle 1: Offline-Evaluationsergebnisse der beiden Faktormodelle. Außer der Anzahl an Faktoren wurden beide mit gleichen Parametern trainiert ($\lambda = 0{,}001$; 16 Iterationen).

	Faktoren	RMSE	NDCG@10
Faktormodell A	10	0,8605	0,8208
Faktormodell B	20	0,8599	0,8223

ergänzt, indem die Webschnittstelle der *The Movie Database*[3] (TMDB) verwendet wurde. Die Entscheidung für die Film-Domäne trafen wir aus zwei zentralen Gründen. Zum einen existieren qualitativ sehr hochwertige Datensätze mit Filmbewertungen, die sich ideal für die Generierung von Empfehlungen verwenden lassen. So muss nicht erst eine aufwendige und kostspielige Erhebung von Bewertungen erfolgen, bevor ein Faktormodell berechnet werden kann. Der zweite entscheidende Vorteil der Verwendung von Filmen als Datengundlage für *LittleMissFits* ist, dass Kinofilmen sehr populär sind und somit potentiell eine sehr große Spielerschaft adressiert werden kann.

Faktormodelle und -kandidaten

Mit Hilfe der Bewertungen wurden zwei unterschiedliche Faktormodelle trainiert. Dazu wurde die *Mahout*-Bibliothek[4] und die Implementierung des *ParallelSGDFactorizer* eingesetzt. Als Parameter verwendeten wir bei beiden Faktorisierungen 0,001 für den Regulierungsparameter λ und trainierten das Modell über 16 Iterationen. Die Modelle unterschieden sich jedoch hinsichtlich der Anzahl ihrer Faktoren. Ein Modell wurde mit 20, das andere mit 10 Faktoren trainiert. Offline-Evaluierung ergab eine gute Vorhersagegenauigkeit mit leichtem Vorteil zugunsten der Faktorisierung mit 20 Faktoren (Tabelle 1).

Für jeden der aus den Faktormodellen stammenden Faktoren wurden 25 Kandidaten ausgewählt (für beispielhafte Kandidaten, siehe Tabelle 2). Wir folgen dabei dem Ansatz von Loepp et al. [14], um repräsentative Filme für die latenten Faktoren zu identifizieren. Bei dieser Methode, wird für jede Film-Faktor-Kombination ein Gewicht berechnet, wie relevant der Film für den entsprechenden Faktor ist. Dabei wird sichergestellt, dass der Film besonders hohe Werte für diesen Faktor, aber neutrale Werte für alle anderen Faktoren hat. Somit ist gewährleistet, dass Kandidaten eines Faktors speziell für diesen Faktor und nicht für andere Faktoren relevant sind. Zudem begünstigt ein Popularitätswert bekannte Filme, um die Wahrscheinlichkeit zu erhöhen, dass die Kandidaten eines Faktors den Spielern bekannt sind. Dieser Popularitätswert wird anhand der Anzahl der abgegebenen Bewertungen für diesen Film ermittelt.

[3]https://www.themoviedb.org/

[4]https://mahout.apache.org/

Tabelle 2: Jeweils drei beispielhafte Kandidaten für die Faktoren aus Faktormodell A.

Faktor	Beispielhafte Kandidaten
1	*Star Wars*, *Babylon 5*, *Time Bandits*
2	*The Guradian*, *Cheech & Chong*, *Beverly Hills Cop*
3	*Spirited Away*, *Pan's Labyrinth*, *Open Hearts*
4	*Blade*, *Starship Troopers*, *The Mummy*
5	*Braveheart*, *Forrest Gump*, *Toy Story*
6	*Back to the Future*, *Ghostbusters*, *Batman*
7	*The Net*, *Ocean's Eleven*, *Ocean's Twelve*
8	*Pulp Fiction*, *Fight Club*, *Trainspotting*
9	*The Lion King*, *Aladdin*, *Titanic*
10	*Gattaca*, *Truman Show*, *I Love You Phillip Morris*

4 NUTZERSTUDIE

Um Spielspaß beim Spielen von *LittleMissFits* und Qualität der mit Hilfe des Spiels erhobenen Daten zu evaluieren, ließen wir das Spiel durch Probanden einer Online-Studie spielen. Dabei wurden insgesamt 570 Runden von 46 Spielern gespielt. Die gespielten Runden verteilten sich zufällig auf die beiden zuvor vorgestellten Faktorisierungen (Tabelle 5). Spieler wurden zudem aufgefordert an einer Umfrage zu empfundenem Spielspaß, Schwierigkeitsgrad und Bekanntheit der Filme teilzunehmen. Von den 46 Personen, die *LittleMissFits* gespielt haben, nahmen 13 (7 weiblich) im Alter von durchschnittlich 27 ($\sigma = 12{,}32$) Jahren an der Umfrage teil.

Game Experience

Die Spieler wurden gebeten mindestens ein Spiel zu spielen. Dennoch spielte im Durchschnitt jeder Spieler 2,5 Spiele[5]. Dies, und das auf einer 5-stufigen *Kunin-Skala* [9] erhobene Fragebogen-Item zum empfunden Spielspaß (($M = 4{,}69$, $\sigma = 1{,}45$)), deutet darauf hin, dass die Spieler Spaß an dem Spiel empfunden haben. Teilweise wurde das Spiel als recht schwer wahrgenommen. Ein entsprechendes Item des Fragebogens ergab, dass 50,0% der Teilnehmer das Spiel als *„schwer"* empfanden, während 41,7% den Schwierigkeitsgrad mit *„Okay"* bewerteten. Auch die Bekanntheit der Filme lag im mittleren Bereich. Während 46,2% angaben die Filme *„ein bisschen"* zu kennen, gaben 23,1% der Spieler an sich *„gut"* mit den im Rahmen des Spiels präsentierten Filmen auszukennen.

Spieldaten

Die Auswertung der aufgezeichneten Spieldaten ergab, dass pro Spieler durchschnittlich 12,39 Runden gespielt wurden.

[5]Die Anzahl Spieler wurde anhand deren Benutzernamen ermittelt. Da nicht ausgeschlossen werden kann, dass sich einzelne Spieler mit einem anderen Benutzernamen erneut auf der Spielseite anmeldeten, liegt die tatsächliche Anzahl gespielter Spiele pro Nutzer wahrscheinlich etwas höher.

Da jede gespielte Runde genau einen Datenwert für die Auswertung produziert, entspricht dieser Wert gleichzeitig der *Expected Contribution* [24] von *LittleMissFits*, welche beschreibt, mit wie viel zusätzlichen Daten pro neuem Spieler gerechnet werden kann. Im Folgenden werden einige Ergebnisse detaillierter diskutiert. Diese beziehen sich, falls nicht anders vermerkt, auf die Faktorisierung, die mit 10 Faktoren trainiert wurde.

Erfolgsquoten der Faktoren. Tabelle 3 zeigt die Erfolgsquoten für jeden der Faktoren, wenn dieser als *Faktor A* verwendet wurde, also als der Faktor, von dem vier der fünf Filme stammten. Als Erfolg wird gewertet, wenn Spieler den fünften Film, also den aus *Faktor B*, korrekt identifizierten. Faktoren unterschieden sich hinsichtlich ihrer Erfolgsquote teilweise stark. Während bei einigen Faktoren (z.B. bei Faktor 4, 9 und 10) die Spieler relativ hohe Erfolgsquoten erzielten (verglichen mit einer Quote von 20% bei einer zufälligen Auswahl), war dies bei den Faktoren 6 und 7 nicht der Fall.

Tabelle 3: Erfolgsquoten der Runden, bei den der jeweilige Faktor als *Faktor A* verwendet wurde, also vier der fünf Filme von diesem Faktor stammten.

Faktor	1	2	3	4	5	6	7	8	9	10
Quote	38%	40%	29%	66%	37%	12%	17%	40%	44%	43%

Erfolgsquoten bei Faktorkombinationen. Um Aussagen über wechselseitige Beziehungen zwischen den Faktoren treffen zu können, z.B. ob diese sich sehr ähnlich sind, werden in Tabelle 4 Ergebnisse zu Runden in Abhängigkeit ihrer beteiligten Faktoren vorgestellt. Die Erfolgsquoten beziehen dabei alle Runden ein, in denen die beiden Faktoren beteiligt sind (also ungeachtet dessen ob als *Faktor A* oder *Faktor B*). Erneut ergibt das Spiel sehr unterschiedliche Werte. So führten manche Kombinationen von Faktoren wesentlich häufiger zu Erfolgen (z.B. die Faktoren 4 und 6) als andere (z.B. Faktoren 1 und 2).

Vergleich der Faktorisierungen. Eines der von uns adressierten Einsatzszenarios für *LittleMissFits* ist der Vergleich unterschiedlich parametrisierter Faktorisierungen. In der hier vorgestellten Studie wurden zwei Faktorisierungen eingesetzt (mit 10 und mit 20 Faktoren). Absolute und relative Erfolge können in Tabelle 5 nachvollzogen werden. Hierbei konnte ein Unterschied bezüglich der Erfolgsquoten, die mit *LittleMissFits* für die beiden Faktorisierungen erzielt wurden, gefunden werden. Konkret zeigte sich, dass die Faktorisierung, die mit 10 Faktoren trainiert wurde, insgesamt zu einer höheren Erfolgsquote führte, verglichen mit der Faktorisierung, welche aus 20 Faktoren bestand.

Tabelle 4: Erfolgsquoten von paarweise kombinierten Faktoren. Bei der mit * markierten Kombination liegen nicht genügend Daten für eine Auswertung vor. Der untere Teil der Tabelle ist symmetrisch zu dem oberen und wurde daher leer gelassen.

	1	2	3	4	5	6	7	8	9	10
1	-	11%	20%	36%	33%	30%	57%	71%	42%	50%
2		-	33%	60%	20%	10%	60%	25%	100%	17%
3			-	20%	38%	*	17%	20%	33%	50%
4				-	71%	88%	50%	67%	58%	40%
5					-	14%	57%	17%	43%	63%
6						-	13%	14%	0%	20%
7							-	33%	50%	0%
8								-	22%	100%
9									-	50%
10										-

Tabelle 5: Verteilung der Antworten und der daraus resultierenden Erfolgsquote bei den unterschiedlich konfigurierten Faktormodellen.

	Antworten		
	korrekt	falsch	Erfolgsquote
Faktormodell A	120	201	37,4%
Faktormodell B	71	178	28,5%
Gesamt	191	379	33,6%

Diskussion

Die mit Hilfe von *LittleMissFits* erhobenen Daten gestatten Einsicht in Eigenschaften von latenten Faktormodellen, die ohne den Einsatz unseres GWAP verborgen geblieben wären. Im Folgenden werden mit Hilfe der Ergebnisse, unsere in der Einleitung aufgeführten Forschungsfragen diskutiert.

FF1: Konsistenz der Faktoren. Die Erfolgsquoten pro Faktor aus Tabelle 3 lassen Rückschlüsse auf deren wahrgenommene Konsistenz zu. Wird eine Runde gespielte ist es notwendig, dass die von dem *Faktor A* stammenden Filme als einheitlich wahrgenommen werden, um jenen Film zu identifizieren, der von *Faktor B* stammt. Es kann daher gefolgert werden, dass die Faktoren mit einer vergleichsweise hohen Erfolgsquote (z.B. Faktor 4) eine sehr klare Semantik transportieren, die von den Probanden einfach zu verstehen ist und als konsistent wahrgenommen wurde. Scheinbar ist Faktor 4 gut verständlich für die Nutzer und könnte daher ein vielversprechender Kandidat sein, um Empfehlungen zu erklären. Im Gegensatz dazu führten andere Faktoren eher selten zu erfolgreichen Runden (z.B. Faktor 6). Diese Faktoren scheinen nicht über eine einfach wahrzunehmende, einheitliche Semantik zu verfügen. Eine Erkenntnis aus diesen Ergebnissen

könnte daher sein, solche Faktoren nicht für die Erklärung der Empfehlungen heranzuziehen.

FF2: Unterscheidbarkeit der Faktoren. Eine weitere Eigenschaft von latenten Faktormodellen, die mit Hilfe von *LittleMissFits* untersucht werden kann ist, wie unterscheidbar die Faktoren untereinander sind. Sind sich zwei Faktoren sehr ähnlich, werden Runden, an denen diese Faktoren beteiligt sind, schwerer zu spielen sein. Die Ergebnisse aus Tabelle 4 können daher Aufschluss darüber geben, ob es Faktoren im Modell gibt, die als leicht zu unterscheiden wahrgenommen werden. Dies scheint z.B. für die Faktoren 4 und 6 der Fall zu sein.

FF3: Vergleich unterschiedlicher Faktormodelle. Werden die Ergebnisse sämtlicher Faktoren eines Faktormodells zusammengefasst, können unterschiedlich parametrisierte Faktorisierungen verglichen und generelle Aussagen über deren inhaltliche Verständlichkeit gemacht werden. Dies ist insbesondere von Interesse, wenn eine Vorauswahl getroffen und eine Faktorisierung ausgewählt werden soll, die sich besonders für die semantische Auswertung der Faktoren eignet. Dies ist etwa der Fall, wenn latente Faktoren mit textuellen Erklärungen verbunden werden, wie in [12]. Solche Ansätze könnten entscheidend davon profitieren ein Faktormodell zu wählen, das eine möglichst hohe intuitive Verständlichkeit besitzt. Nach der Auswertung unserer Ergebnisse, scheint insbesondere eine geringe Faktoranzahl für deren inhaltliche Klarheit zuträglich zu sein.

Diskussion der Game Experience. Für ein GWAP ist es wichtig, dass die Spieler Spaß an dem Spiel empfinden. Durch die Fragebogen-Items konnte festgestellt werden, dass *LittleMissFits* bereits relativ hohe Werte bezüglich des Spielspaßes erzielt. Wir führen dies unter Anderem auf die sehr spielerischen Grafiken und den narrativen Hintergrund zurück, welcher von den Spielern angenommen worden zu sein scheint. So bezogen sich Probanden auf die Grafiken und adaptierten die Narration, als sie im Fragebogen Freitextkommentare hinterließen wie z.B. *„Die Fledermaus ist echt süß."* und *„Manchmal verstehe ich die Fledermaus nicht."*. Dies weist in die selbe Richtung, wie bisherige Erkenntnisse zur positiven Wirkung von Narration auf empfundenen Spielspaß in GWAP [17] und wir leiten daraus die Empfehlung ab, dass sich Entwickler zukünftiger GWAP nicht allein auf typische Spielelemente wie Punkte und Bestenlisten konzentrieren sollten, sondern auch auf visuelle Aspekte und das narrative Setting in dem das Spiel eingebettet wird.

Trotz des hoch bewerteten Spielspaßes, besteht Verbesserungspotential in der Wiederspielrate von *LittleMissFits*. Hier könnte sich negativ ausgewirkt haben, dass das Spiel als recht schwierig zu spielen empfunden wurde. Ein Grund für die Schwierigkeit des Spiels ist sicherlich die teilweise mangelnde Bekanntheit der dargestellten Filme. Dies wurde auch durch Freitextkommentare der Probanden gestützt. Wahrscheinlich ist die teilweise mangelnde Bekanntheit der Filme auf die Bestimmung der Kandidaten für die Faktoren zurückzuführen. Wie zuvor beschrieben, geht in die Bestimmung der Kandidaten jedes Faktors, neben der Relevanz eines Films auch dessen Popularität ein (Anzahl bisher vergebener Bewertungen). Während dies sehr erfolgreiche Filme begünstigt, werden auch solche Filme hervorgehoben, die aufgrund der Tatsache, dass sie schon lange in der Datenbank vorhanden sind, viele Bewertungen erhalten haben. Hier könnte die Aktualität der Filme als Gegengewicht Berücksichtigung finden. Auch könnten Faktoren, die zu insgesamt geringen Erfolgsquoten führen, ab dem Vorhandensein einer bestimmten Menge an Daten nicht weiter vom Spiel berücksichtigt werden, um so einer negativen Auswirkung auf den wahrgenommenen Schwierigkeitsgrad entgegenzuwirken.

5 FAZIT UND AUSBLICK

In diesem Beitrag stellen wir *LittleMissFits* vor, ein Game-with-a-Purpose zur Messung semantischer Konsistenz in latenten Faktoren. Die bislang erhobenen Daten deuten darauf hin, dass das Spiel gut geeignet ist, um Konsistenz innerhalb der Faktoren und Diversität zwischen den Faktoren zu messen. Das Spiel wurde bislang mit zwei Faktorisierungen mit unterschiedlichen Anzahlen der Faktoren getestet, soll in Zukunft aber noch auf weitere Parametrisierungen der MF übertragen werden. Obwohl sicherlich auch weitere Werte für die Faktoranzahl mit dem Spiel evaluiert werden sollten, stehen insbesondere noch Versuche mit anderen λ–Werten aus. Dieser Regulierungsfaktor könnte einen großen Einfluss auf die Verständlichkeit der Faktoren haben, da mit ihm die Lernrate während der Trainingsphase reguliert wird, um eine Überanpassung der Faktoren zu verhindern. Eventuell eignen sich jedoch genau solche „überangepassten" Faktoren gut dazu, um die Empfehlungen zu erklären. Zudem wird eine wesentlich größere Anzahl an Spielern benötigt, weshalb solche Versuche bislang noch ausstehen. Um eine noch stärkere Motivation der Spieler zu erzeugen, planen wir eine tiefer greifende Implementierung von Gamification, z.B. in Form von zusätzlichen kollaborativen Spielelementen. Auch wäre eine Übertragung auf weitere Hintergrunddomänen denkbar. So könnten z.B. Inhalte von Lehrveranstaltungen, wie etwa einer Vorlesungsreihe, bezüglich ihrer semantischen Übereinstimmung durch die Spieler ausgewertet werden. Hieraus ließen sich möglicherweise ganz neue Formen der Kategorisierung von Lerninhalten identifizieren, welche zu einer harmonischeren Komposition von Lehrveranstaltungen führen könnte.

LITERATUR

[1] Sam Banks, Rachael Rafter, und Barry Smyth. 2015. The Recommendation Game: Using a Game-with-a-Purpose to Generate Recommendation Data. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. ACM, New York, NY, USA, 305–308. https://doi.org/10.1145/2792838.2799675

[2] Emden Gansner, Yifan Hu, Stephen Kobourov, und Chris Volinsky. 2009. Putting Recommendations on the Map - Visualizing Clusters and Relations. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*. ACM, New York, NY, USA, 345–348. https://doi.org/10.1145/1639714.1639784

[3] Carlos A. Gomez-Uribe und Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems* 6, 4 (2015), 13:1–13:19. https://doi.org/10.1145/2843948

[4] Severin Hacker und Luis von Ahn. 2009. Matchin: Eliciting User Preferences with an Online Game. In *Proceedings of the 27th ACM International Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1207–1216. https://doi.org/10.1145/1518701.1518882

[5] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, und Kuan-Ta Chen. 2009. KissKissBan: A Competitive Human Computation Game for Image Annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, New York, NY, USA, 11–14. https://doi.org/10.1145/1600150.1600153

[6] Bart P. Knijnenburg und Martijn C. Willemsen. 2015. Evaluating Recommender Systems with User Experience. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, und Bracha Shapira (Hrsg.). Springer US, Boston, MA, 309–352. https://doi.org/10.1007/978-1-4899-7637-6{_}9

[7] Joseph A. Konstan und John Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[8] Yehuda Koren, Robert M. Bell, und Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[9] Theodore Kunin. 1955. The Construction of a New Type of Attitude Measure1. *Personnel Psychology* 8, 1 (1955), 65–77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x

[10] Johannes Kunkel, Benedikt Loepp, und Jürgen Ziegler. 2017. A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 3–15. https://doi.org/10.1145/3025171.3025189

[11] Johannes Kunkel, Benedikt Loepp, und Jürgen Ziegler. 2018. Ein Online-Spiel zur Benennung latenter Faktoren in Empfehlungssystemen. In *Mensch und Computer 2018 - Tagungsband*, Raimund Dachselt und Gerhard Weber (Hrsg.). Gesellschaft für Informatik e.V, Bonn. https://doi.org/10.18420/muc2018-mci-0108

[12] Johannes Kunkel, Benedikt Loepp, und Jürgen Ziegler. 2018. Understanding Latent Factors Using a GWAP. In *Proceedings of the Late-Breaking Results track part of the Twelfth ACM Conference on Recommender Systems (RecSys'18)*. https://arxiv.org/pdf/1808.10260.pdf

[13] Benedikt Loepp, Tim Donkers, Timm Kleemann, und Jürgen Ziegler. 2019. Interactive recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies* (2019), 21–41. Issue 121. https://doi.org/10.1016/j.ijhcs.2018.05.002

[14] Benedikt Loepp, Tim Hussein, und Jürgen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd International Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3085–3094. https://doi.org/10.1145/2556288.2557069

[15] Lennart E. Nacke, Chris Bateman, und Regan L. Mandryk. 2014. BrainHex: A neurobiological gamer typology survey. *Entertainment Computing* 5, 1 (2014), 55–62. https://doi.org/10.1016/j.entcom.2013.06.002

[16] B. Németh, G. Takács, I. Pilászy, und D. Tikk. 2013. Visualization of movie features in collaborative filtering. In *2013 IEEE 12th International Conference on Intelligent Software Methodologies, Tools and Techniques (SoMeT)*. 229–233. https://doi.org/10.1109/SoMeT.2013.6645674

[17] Nathan R. Prestopnik und Jian Tang. 2015. Points, stories, worlds, and diegesis: Comparing player experiences in two citizen science games. *Computers in Human Behavior* 52 (2015), 492–506. https://doi.org/10.1016/j.chb.2015.05.051

[18] M. Rossetti, F. Stella, und M. Zanker. 2013. Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems. In *2013 24th International Workshop on Database and Expert Systems Applications*. IEEE Computer Society, 162–167. https://doi.org/10.1109/DEXA.2013.26

[19] Nitin Seemakurty, Jonathan Chu, Luis von Ahn, und Anthony Tomasic. 2010. Word Sense Disambiguation via Human Computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 60–63. https://doi.org/10.1145/1837885.1837905

[20] Brent Smith und Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *Internet Computing, IEEE* 21, 3 (2017), 12–18. https://doi.org/10.1109/MIC.2017.72

[21] Barry Smyth, Rachael Rafter, und Sam Banks. 2016. Harnessing Crowdsourced Recommendation Preference Data from Casual Gameplay. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2930238.2930260

[22] Nava Tintarev und Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, und Bracha Shapira (Hrsg.). Springer US, Boston, MA, 353–382. https://doi.org/10.1007/978-1-4899-7637-6_10

[23] Luis von Ahn. 2006. Games with a Purpose. *Computer* 39, 6 (2006), 92–94. https://doi.org/10.1109/MC.2006.196

[24] Luis von Ahn und Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM* 51, 8 (2008), 58–67. https://doi.org/10.1145/1378704.1378719

[25] Greg Walsh und Jennifer Golbeck. 2010. Curator: A Game with a Purpose for Collection Recommendation. In *Proceedings of the 28th ACM International Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2079–2082. https://doi.org/10.1145/1753326.1753643