

CAVE-basierte Nutzerstudie eines Aufenthalts-Empfehlungssystems

Jens Hegenberg¹, Niclas Kannengiesser¹, Elisa Klose¹, Ludger Schmidt¹

Fachgebiet Mensch-Maschine-Systemtechnik, Universität Kassel¹

J.Hegenberg@uni-kassel.de, N.Kannengiesser@uni-kassel.de,
E.Klose@uni-kassel.de, L.Schmidt@uni-kassel.de

Zusammenfassung

Reiseassistenzsysteme sind vielversprechend, um individualisierte und kontextbezogene Unterstützung entlang intermodaler Reiseketten im öffentlichen Verkehr anzubieten. Hierbei stellen sowohl zeitkritische Situationen als auch solche mit längerer Wartezeit eine Belastung des Reisenden dar. Für letztere bietet das untersuchte System Assistenz durch individualisierte Gestaltungsvorschläge für längere Aufenthalte. In Vorbereitung auf eine großangelegte Feldstudie wurden vier Varianten des zugrundeliegenden Empfehlungssystems und ein Trainingsansatz zur Initialisierung des Nutzerprofils hinsichtlich der numerischen und der subjektiv wahrgenommenen Genauigkeit der Empfehlungen evaluiert. Die Benutzungsschnittstelle in einer optischen Durchsichtdatenbrille wird außerdem hinsichtlich ihrer Gebrauchstauglichkeit evaluiert. Für die Evaluation wurde eine Reise inkl. zweier längerer Aufenthalte mittels einer CAVE simuliert.

1 Einleitung

Im Reisekontext finden Assistenzsysteme in Form von Smartphone-Apps zunehmende Verbreitung. Diese bieten in der Regel Echtzeitinformationen zur geplanten Mobilitätskette und häufig auch Augmented-Reality-(AR-)Visualisierungen mit Video-Durchsicht, die bspw. Position und Entfernung nahegelegener Haltestellen darstellen. In dem Forschungsprojekt, das den Rahmen dieses Beitrags bildet, stehen AR-Visualisierungen in optischen Durchsichtdatenbrillen im Fokus, die bspw. die Navigation durch komplexe Mobilitätsknoten, z. B. Bahnhöfe oder Flughäfen, unterstützen können. Zur Lokalisierung im Innenraum sind erste Datenbrillen mittlerweile mit Sensorik zur Umgebungserfassung für den Nahbereich ausgestattet. Zusätzlich wird von den Betreibern großer Mobilitätsknoten zunehmend der Einsatz von Bluetooth-Baken erprobt, die eine globale Lokalisierung ermöglichen.

Die Integration adaptiver und lernender Assistenzfunktionen ermöglicht weiterhin die Adaptierung an den Nutzungskontext und die Bedürfnisse des Nutzers. Letztere werden dazu über die Nutzungsdauer des Systems gelernt. Dieser Beitrag konzentriert sich auf eine Assistenzfunktion, die an einem Mobilitätsknoten Möglichkeiten zur Gestaltung längerer Aufenthalte aufzeigt, die zum Kontext, bspw. der Aufenthaltsdauer, und zu den Interessen des Nutzers passen. Die Vorschläge werden mittels eines Empfehlungssystems generiert und können bspw. den Besuch eines Restaurants oder eines Geschäfts beinhalten.

2 Empfehlungssysteme

Empfehlungssysteme sind bislang vor allem im E-Commerce verbreitet und werden dort genutzt, um Kunden individuell passende Produkte vorzuschlagen (Konstan & Riedl, 2012). Auch in Anwendungsbereichen, wie bspw. Musik- oder Videostreaming, werden Empfehlungssysteme intensiv genutzt. Alle genannten Anwendungsbereiche profitieren davon, dass sie vorwiegend über Computersysteme genutzt werden, wodurch eine umfangreiche Datenerhebung zum Nutzerverhalten ermöglicht wird. Bspw. ermöglicht die sekundengenaue Erfassung, wann welcher Musiktitel wie lange gehört wurde, dass Schlussfolgerungen zum Musikgeschmack der Nutzer auf großen Datenmengen basieren können. Im Reisekontext sind Empfehlungssysteme noch wenig verbreitet, und die Erhebung des Nutzerverhaltens im realen Kontext ist deutlich komplexer. So ist es bspw. in der Regel nicht automatisch möglich zu ermitteln, welches Gericht in einem Restaurant bestellt wurde.

Der grundlegende Aufbau eines Empfehlungssystems ist in Abbildung 1 schematisch dargestellt. Empfehlungssysteme wählen aus einer Menge sogenannter Items diejenigen aus, die für den Nutzer im aktuellen Nutzungskontext relevant sind (Klahold, 2009; Ricci et al., 2011). Die gewählten Items stellen die Empfehlungen dar. Die betrachteten Items sind dabei beliebig, bspw. können dies Gegenstände, Orte oder Filme sein (Ricci et al., 2011). Die Repräsentation der Items innerhalb des Empfehlungssystems erfolgt entsprechend einer vorgegebenen Datenstruktur, dem Item-Modell. Analog werden der Nutzungskontext entsprechend dem Kontext-Modell und die Nutzereigenschaften entsprechend dem Nutzer-Modell repräsentiert. Eine konkrete Ausprägung des Kontexts wird als Setting und eine konkrete Ausprägung der Nutzereigenschaften als Nutzerprofil bezeichnet.

Das initiale Datenset ist der Datenbestand bzgl. Items und Nutzerprofil bei Inbetriebnahme des Empfehlungssystems. Zu diesem Zeitpunkt liegen in der Regel kaum Daten über den Nutzer vor, so dass darauf basierende Empfehlungen nicht möglich sind. Diese Problematik wird als Kaltstartproblem bezeichnet (Schafer et al., 2007; Lika et al., 2014). Ein Lösungsansatz besteht in einer initialen vorgeschalteten Trainingsphase, in der der Nutzer mittels paarweiser Vergleiche repräsentativer Items seine Präferenzen initialisiert (Hüllermeier et al., 2008).

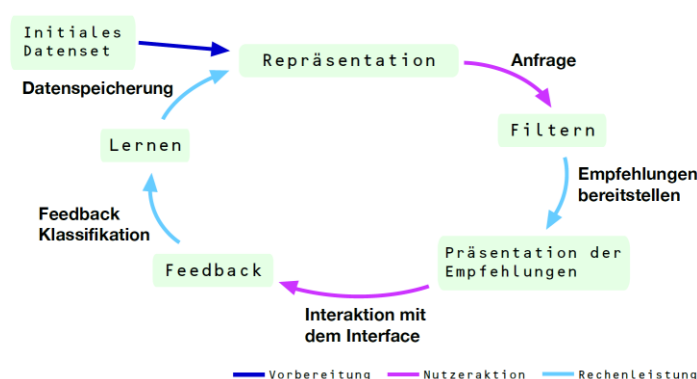


Abbildung 1: Schematischer Aufbau eines lernenden Empfehlungssystems

Nutzeranfragen an das Empfehlungssystem können explizit, bspw. mittels einer Benutzerschnittstelle, oder implizit, bspw. durch Ankunft an einem Bahnhof, gestellt werden. Bei einer Anfrage wird für jedes Item eine Bewertung, der (Item-)Score, berechnet, der die Relevanz des Items für das aktuelle Setting und das aktuelle Nutzerprofil widerspiegelt. Die Berechnung der Scores wird als Filtern und die verschiedenen Berechnungsverfahren als Filter bezeichnet (Burke, 2007). Die Auswahl des passenden Filters richtet sich nach den zur Verfügung stehenden Daten und nach der vorgesehenen Nutzerinteraktion mit dem System.

Der Filterprozess lässt sich in drei serielle Filterstufen gliedern (Burke, 2007). Der Content-Based-Filter berechnet im ersten Schritt die Item-Scores derart, dass sie die Passung zwischen Item und Nutzerprofil widerspiegeln. Anschließend sortiert der Constraint-Based-Filter alle Items aus, die gegen definierbare Randbedingungen verstoßen, bspw. Ausschluss von Raucherkeipen für Nichtraucher. Am Ende des Filterprozesses passt der kontextbezogene Filter (Adomavicius et al., 2011) die Empfehlungen an das aktuelle Setting an. Dies berücksichtigt bspw. Öffnungszeiten oder bereits genutzte Items.

Zur Berechnung der Scores im Content-Based-Filter können erweiterte Vector-Space-Modelle (VSM) genutzt werden (Musto, 2010). In einem VSM werden sowohl Items als auch das Nutzerprofil durch Vektoren repräsentiert. Den Dimensionen der Vektoren werden Attribute zugeordnet. Ist das Attribut, bspw. die Verfügbarkeit von freiem WLAN, für ein Item zutreffend, wird im Item-Vektor eine Eins, sonst eine Null eingetragen. Das Nutzerprofil wird ebenfalls als ein solcher Attributvektor repräsentiert, wobei die Dimensionen des Item-Modells und des Nutzer-Modells identischen Attributen zugeordnet werden. Während die Dimensionen des Item-Modells mit Werten aus der Menge $\{0, 1\}$ belegt werden, stammen die Werte der Nutzerprofils aus dem reellen Intervall $[-1, 1]$. Minus Eins entspricht dabei einer negativen Bewertung des Attributs, Null einer neutralen Bewertung und Eins einer positiven Bewertung. Die Ähnlichkeit zwischen solchen Item-Vektoren untereinander und zwischen einem Item-Vektor und dem Nutzerprofil-Vektor wird mittels der Kosinus-Ähnlichkeit berechnet.

Der Filterprozess kann um zwei optionale Komponenten erweitert werden. Erstens ist dies die Gewichtung häufig vorkommender Attribute mittels des Tf-IDF-Maßes, um zu vermei-

den, dass Attribute lediglich aufgrund ihrer Häufigkeit starken Einfluss auf die Scores haben. Die Gewichtung $w \in [0; 1]$ der einzelnen Attribute berechnet sich nach Formel 1 und wird beim Content-Based-Filter angewendet.

$$w(t, d) = tf_{t,d} * IDF_t \text{ mit } IDF_t = \log\left(\frac{|D|}{1 + |\{d \in D \mid t \in d\}|}\right)$$

Formel 1: Berechnung der Gewichtung w auf Basis der Vorkommenshäufigkeit (tf , Term-Frequency) des Attributs t im Item d und der inversen Dokumenthäufigkeit (IDF , Inverse-Document-Frequency), wobei D der Menge aller bewerteten Item entspricht

Zweitens ist dies ein Serendipitäts-Mechanismus (Kotkov et al., 2016), der nach dem Constraint-Based-Filter greift und gezielt unerwartete Items in die Empfehlungen einstreut. Dadurch können u. a. neue Anreize geschaffen und somit der Gefahr eines Overfittings entgegengewirkt werden. Dazu werden gezielt k -viele nicht genutzte Items ausgewählt, deren Ähnlichkeiten zu dem Item mit dem höchsten Score minimal sind (k -Furthest-Neighbours). Die Anzahl k ist auf maximal 10 % der Items begrenzt, die den Constraint-Based-Filter passiert haben. Den ausgewählten Items wird ein zufälliger Score zwischen ihrem tatsächlichen und dem höchsten Score zugeordnet.

Die Antwort auf eine Anfrage kann dem Nutzer als Liste der N Empfehlungen mit dem höchsten Item-Score, der Top- N -Item-Liste, präsentiert werden, mit der der Nutzer, bspw. durch Auswahl eines Items, interagieren kann (Zhao et al., 2017). Die Interaktion kann explizit, bspw. durch Bewertung einer Empfehlung, oder implizit, bspw. durch die Nutzung empfohlener Items, erfolgen und dient als Feedback zur Aktualisierung des Nutzerprofils. Diese Aktualisierung kann als maschineller Lernprozess beschrieben werden, der nach und nach die Interessen des Nutzers lernt. Lernende System können adaptierend oder adaptiv sein (Oppermann, 1994), wobei adaptierende Funktionen auf der manuellen Veränderung von Systemparametern basieren und adaptive Funktionen das System automatisch z. B. an Setting und Nutzerprofil anpassen.

Bei der Evaluation von Empfehlungssystemen können u. a. die Kriterien Genauigkeit der Empfehlungen und Gebrauchstauglichkeit herangezogen werden (Gunawardana & Shani, 2015).

3 Konzept des Aufenthalts-Empfehlungssystems

Das betrachtete Assistenzsystem zur Gestaltung längerer Aufenthalte an Mobilitätsknoten betrachtet als Items, die an den Mobilitätsknoten einer Reisekette vorzufindenden Angebote verschiedener Domänen, bspw. Gastronomie und Einzelhandel. Das Item-Modell orientiert sich an den POI-Klassifikationen von OpenStreetMaps, vom Deutschen Hotel- und Gaststättenverbands u. a. Die Daten der Items werden aus POI-Datenbanken geladen und entsprechend des Item-Modells aufbereitet.

Das Nutzer-Modell basiert auf einer Recherche zu häufig bei der Reiseassistenz verwendeten Benutzer-Charakteristika und den Ergebnissen eigener Fokusgruppenuntersuchungen. Das Nutzerprofil umfasst sowohl explizit anzugebende, bspw. Raucher oder Nichtraucher, und automatisch erfasste Daten, bspw. die durchschnittliche Laufgeschwindigkeit oder die Nutzerinteressen.

Das Kontext-Modell basiert auf Literaturrecherchen zu Umgebungsfaktoren beim Reisen und den Ergebnissen eigener Fokusgruppenuntersuchungen. Die ermittelten Faktoren wurden hinsichtlich ihrer Relevanz und praktischen Erfassbarkeit ausgewählt. Während der Reise wird das Setting bspw. durch Echtzeit-Informationen zu Verspätungen im öffentlichen Verkehr aktualisiert.

Zur Lösung des Kaltstartproblems wird bei Inbetriebnahme des Reiseassistenzsystems eine Trainingsphase vorgeschaltet, die mittels paarweisen Vergleichs acht repräsentativer Items zur Initialisierung des Nutzerprofils dient. Der Nutzer muss dabei die jeweils präferierte Aufenthaltsmöglichkeit auswählen.

Sobald die Aufenthaltsdauer an einem Mobilitätsknoten einen konfigurierbaren Grenzwert übersteigt, wird eine Anfrage an das Empfehlungssystem ausgelöst. Der Content-Based-Filter berechnet daraufhin Item-Scores mittels des VSM-Ansatzes. Der Constraint-Based-Filter sortiert anschließend alle Items, die nicht dem aktuellen Verkehrsknoten zugeordnet sind oder aufgrund von Altersbeschränkungen, bspw. bei Raucher kneipen, unzulässig sind, aus. Der kontextbezogene Filter berücksichtigt abschließend das aktuelle Setting, indem zunächst die Erreichbarkeit eines Items unter Berücksichtigung von Wegzeiten zum Item und vom Item zum Abfahrtsort sowie der Zugänglichkeit geprüft wird. Im nächsten Schritt werden die Scores bereits genutzter und ähnlicher Items abgewertet. Dabei wurden starke Vereinfachungen bzgl. der zeitlichen Entwicklung der Nutzerbedürfnisse getroffen. Genutzte Items selbst werden nach der Nutzung zunächst nicht mehr empfohlen. Weiter wird angenommen, dass auch ähnliche Items temporär nicht empfohlen werden sollten. Der Nutzungszweck wird dabei vernachlässigt. Abwertung und Regeneration der Scores verschiedener Domänen werden als gleich angenommen. Diese Annahme könnte bei gastronomischen Angeboten zutreffen, ist beim Einzelhandel allerdings kritisch zu hinterfragen. Während der Nutzung wird der Score auf -1 abgewertet und regeneriert dann mit fortschreitender Zeit entsprechend der Formel 2. Den genutzten Items ähnliche Items werden entsprechend der Formel 3 abgewertet. Zusätzlich sollen Filtervarianten mit einer Tf-IDF-Gewichtung und/oder einem Serendipitäts-Mechanismus umgesetzt werden.

$$s_{tmp}(it) = \begin{cases} s - \frac{\cos\left(\frac{t - t_{u,it}}{x}\right) + 1}{2} * (s + 1), & 0 \leq \frac{t - t_{u,it}}{x} < \pi \\ s, & \text{sonst} \end{cases}$$

Formel 2: Temporäre Abwertung des Scores s eines Items it , das zum Zeitpunkt $t_{u,it}$ genutzt wurde, mit dem aktuellen Zeitpunkt t und einem Streckfaktor x zur Verlangsamung der Score-Regeneration

$$s_{tmp}(it1) = \begin{cases} s_{it1} - \frac{\cos\left(\frac{t - t_{u,it2}}{x}\right) + 1}{2} * (s_{it1} + 1) * similarity(it1, it2), & 0 \leq \frac{t - t_{u,it2}}{x} < \pi \\ s_{it1}, & \text{sonst} \end{cases}$$

Formel 3: Temporäre Abstufung des Scores s eines Items $it1$ wegen der Nutzung eines ähnlichen Items $it2$ mit dem aktuellen Zeitpunkt t , dem Nutzungszeitpunkt $t_{u,it2}$ von $it2$ und dem Streckfaktor x

Bei der Präsentation der Empfehlungen erscheint eine umfangreiche Top-N-Item-Liste aufgrund ihres Platzbedarfs im Sichtfeld einer Durchsichtdatenbrille nicht praktikabel. Es werden daher immer nur drei Empfehlungen gleichzeitig angezeigt. Feedback erfolgt bspw. durch Auswahl einer Empfehlung zur Nutzung oder indem Details zu einer Empfehlung aufgerufen werden. Auf Basis dieses Feedbacks passt sich das System während der Nutzung adaptiv an das Nutzerverhalten an. Auch eine Adaptierung durch explizite Einstellung von Randbedingungen ist möglich.

4 Prototypische Realisierung

Das technische System umfasst im Wesentlichen drei Komponenten: Einen Desktop-Rechner, der das Empfehlungssystem ausführt und dem Versuchsleiter zur Ablaufsteuerung dient, einer Durchsicht-Datenbrille Microsoft HoloLens zur Visualisierung der Empfehlungen und einer Virtual-Reality-(VR-)Umgebung in einer CAVE zur Simulation der Reisekette und der dabei besuchten Mobilitätsknoten und Aufenthalte (Abbildung 2). Die Sprachsteuerung wurde mittels eines Wizard-of-Oz-Ansatzes realisiert, wobei der Versuchsleiter die Funktionen des Assistenzsystems fernsteuert. Eine Benutzungsschnittstelle für die Nutzerregistrierung und das initiale Training sind ebenfalls integriert.



Abbildung 2: Proband in der CAVE mit Durchsichtdatenbrille an einem virtuellen Mobilitätsknoten.

Die Visualisierung der Top-3-Item-Liste in der Datenbrille zeigt Abbildung 3 links, wobei wegen der guten Lesbarkeit ein Billboard-Style zur Darstellung genutzt wird (Debernardis et al., 2014). Der Benutzer kann mittels frei wählbarer Sprachbefehle einzelne Empfehlungen selektieren (Abbildung 3 mittig), Details zu Empfehlungen anzeigen lassen (Abbildung 3 rechts), weitere Empfehlungen mit niedrigeren Scores anzeigen lassen oder die Navigation zu einem empfohlenen Item und damit dessen Nutzung starten. Erfolgt keine Interaktion bis der grüne Countdown-Balken (Abbildung 3 links) abgelaufen ist, werden die Empfehlungen ausgeblendet und erst nach Nutzerinteraktion wieder angezeigt.

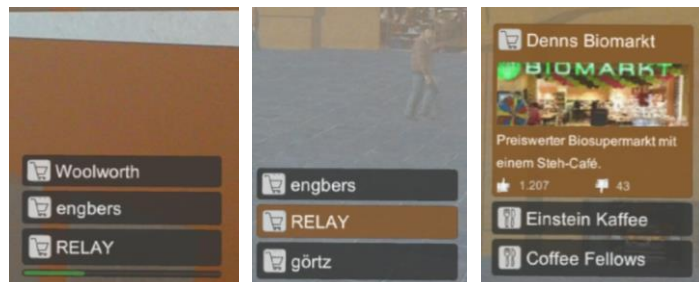


Abbildung 3: Darstellung dreier Empfehlungen mit Countdown-Balken (links), Selektion der zweiten Empfehlung (mittig) und Detailansicht zur ersten Empfehlung (rechts)

Bei der Initialisierung der CAVE werden aus einer Item-Datenbank mit 49 Items 45 Items zufällig ausgewählt und den Ladengeschäften am virtuellen Mobilitätsknoten zufällig zugewiesen. Der Mobilitätsknoten orientiert sich am Frankfurter Hauptbahnhof und wurde mit zufallsgesteuerten Personen und Zügen sowie aufgezeichneten Bahnhofsgerauschen realitätsnah dargestellt. Die Navigation zu den Positionen von Items wird durch die Einblendung eines virtuellen Pfades unterstützt. Die Nutzung eines Items und deren Dauer werden dem Nutzer mittels Texteinblendungen erläutert, also nicht zeitlich simuliert.

5 Nutzerstudie

In Vorbereitung auf eine groß angelegte Feldstudie soll diese CAVE-basierte Nutzerstudie die Auswahl einer Variante des Filterprozesses unterstützen sowie die Gebrauchstauglichkeit der realisierten Visualisierungen und Interaktionsmechanismen sicherstellen. Da keine Testdatensätze für Empfehlungssysteme im Reisekontext verfügbar sind und neben der rechnerischen Empfehlungsgüte auch subjektive Eindrücke betrachtet werden, wurde die Evaluation als Nutzerstudie durchgeführt.

Als unabhängige Variable wurde die Verwendung unterschiedlicher Filterprozesse festgelegt. Variiert werden kann die Aktivierung der Tf-IDF-Gewichtung und / oder des Serendipitäts-Mechanismus. Damit ergeben sich vier Varianten, die in einem Between-Subject-Ansatz evaluiert werden. Zur Bildung homogener Versuchsgruppen (VG) wurde im Vorfeld von allen Probanden ein Fragebogen ausgefüllt, der Alter und Geschlecht sowie die Technikaffinität mittels des TA-EG-Fragebogens (Karrer et al., 2009) erfasste. Anhand dieser Daten

wurden die Probanden ausgewählt und vier möglichst homogene Versuchsgruppen gebildet (vgl. Tabelle 1).

| | | Serendipitäts-Mechanismus | |
|-------------------|------|---------------------------|------|
| | | ohne | mit |
| Tf-IDF-Gewichtung | ohne | VG 1 | VG 2 |
| | mit | VG 3 | VG 4 |

Tabelle 1: Übersicht der Versuchsgruppen

Abhängige Variablen sind die numerische und die vom Nutzer wahrgenommene Genauigkeit der Empfehlungen sowie die Gebrauchstauglichkeit. Die numerische Genauigkeit wurde als Abweichung des Listenindexes der gewählten Empfehlung vom Item mit dem höchsten Score bzw. Listenindex Eins berechnet. Die wahrgenommene Genauigkeit wurde mittels des ResQue-Fragebogens gemessen (Pu et al., 2012) und die Gebrauchstauglichkeit mittels des System-Usability-Scale-(SUS-)Fragebogens (Brooke, 1995). Darüber hinaus wurde qualitatives Feedback der Probanden während des Versuchs und bei einer abschließenden Befragung erfasst.

Die Hypothesen postulieren jeweils eine Abhängigkeit der drei abhängigen Variablen von der genutzten Realisierungsvariante. Die Tf-IDF-Gewichtung sollte zu erhöhter Genauigkeit und damit erhöhter Gebrauchstauglichkeit führen. Der Serendipitäts-Mechanismus sollte entsprechend zu verringerter Genauigkeit und Gebrauchstauglichkeit führen.

5.1 Durchführung

Der Versuchsablauf sah zuerst die Nutzerregistrierung am System unter Angabe demographischer Daten und die Ausführung des Trainings zur Initialisierung des Nutzerprofils vor. Anschließend wurden CAVE, Datenbrille und die virtuelle Reisekette vorgestellt. Dabei wurde deutlich gemacht, dass während des Versuchs die empfohlenen Items genutzt werden sollen und keine Schlagwortsuche möglich ist.

Dann wurde die virtuelle Reise gestartet, die zwei Aufenthalte am virtuellen Mobilitätsknoten umfasste. Das 3-D-Modell war bei beiden Aufenthalten identisch, aber die Auswahl und Zuordnung der Aufenthaltsmöglichkeiten wurde zufällig variiert. Der erste Aufenthalt von 60 Minuten Dauer fand planmäßig statt. Der zweite Aufenthalt wurde anfangs mit 3 Minuten veranschlagt, aber durch eine eintreffende Verspätungsinformation ebenfalls auf eine Gesamtdauer von 60 Minuten verlängert. Der Proband hatte somit in beiden Fällen ausreichend Zeit zur Verfügung, um das Assistenzsystem zur Aufenthaltsgestaltung nutzen zu können.

Bei Ankunft des Nutzers am virtuellen Mobilitätsknoten wurden die ersten drei Empfehlungen in der Datenbrille angezeigt. Nachdem der Nutzer sich mittels Sprachkommandos für die Nutzung einer Empfehlung entschieden hatte, wurde er zur Position des ausgewählten Items navigiert. Bei der Ankunft wurde die virtuelle Nutzungsdauer per Texteinblendung mitgeteilt. In Realzeit wurden direkt anschließend die neuen Empfehlungen präsentiert. Die virtu-

ellen Nutzungszeiten der Aufenthaltsmöglichkeiten wurden so berechnet, dass der Proband während jedes Aufenthalts genau 5 Items nutzte.

Am Versuchsende füllte jeder Proband den ResQue- und den SUS-Fragebogen aus und wurde zu seinen Eindrücken und Verbesserungsvorschlägen befragt.

5.2 Ergebnisse

An der Evaluation nahmen 16 Probanden, davon acht männlich und acht weiblich, im Alter zwischen 20 und 25 Jahren ($M = 24,4$ Jahre; $SD = 2,1$ Jahre) teil. In den Versuchsgruppen waren die Geschlechter gleich verteilt. VG 1 (Alter: $M = 23,8$ Jahre; $SD = 2,9$ Jahre, TA-EG: $M = 3,46$; $SD = 0,13$) nutzte weder die Tf-IDF-Gewichtung noch den Serendipitäts-Mechanismus, VG 2 (Alter: $M = 24,0$ Jahre; $SD = 1,2$ Jahre, TA-EG: $M = 3,08$; $SD = 0,23$) nutzte nur den Serendipitäts-Mechanismus, VG 3 (Alter: $M = 24,8$ Jahre; $SD = 2,4$ Jahre, TA-EG: $M = 3,43$; $SD = 0,31$) nur die Tf-IDF-Gewichtung und für VG 4 (Alter: $M = 25,0$ Jahre; $SD = 1,2$ Jahre, TA-EG: $M = 3,25$; $SD = 0,46$) wurden beide Funktionen aktiviert.

Da die numerische Genauigkeit in den Versuchsgruppen nicht normalverteilt war (Shapiro-Wilk-Test), wurde ein Kruskal-Wallis-Test durchgeführt. Dieser ergab, dass die numerische Genauigkeit für VG 1 signifikant größer als für VG 2 ($p < 0,001$) und VG 3 ($p < 0,01$) ist. Auch die numerische Genauigkeit für VG 4 war signifikant größer als für VG 2 ($p < 0,01$) und VG 3 ($p < 0,01$).

Die Ergebnisse der wahrgenommenen Qualität der Empfehlungen aus dem ResQue-Fragebogen waren in den Versuchsgruppen normalverteilt (Shapiro-Wilk-Test). Mittels ANOVA konnten keine signifikanten Effekte festgestellt werden.

Bezüglich der Gebrauchstauglichkeit ergaben sich für das Assistenzsystem insgesamt SUS-Werte zwischen 62,5 und 95 ($M = 80,5$; $SD = 10,6$). Nach Bangor et al. (2009) liegt dieses Ergebnis deutlich in dem SUS-Intervall, das Systeme beschreibt, die von Nutzern akzeptiert werden (ab einem SUS-Wert von 70). Bei einer Interpretation des Ergebnisses in Schulnoten entspräche die Bewertung der Schulnote „gut“. Dies lässt die Schlussfolgerung zu, dass es sich um ein gebrauchstaugliches System handelt, welches jedoch noch Potential zur Verbesserung der Gebrauchstauglichkeit hat. Für die einzelnen Versuchsgruppen ergaben sich die folgende SUS-Werte (VG 1: $M = 87,5$; $SD = 4,0$, VG 2: $M = 78,8$; $SD = 11,4$, VG 3: $M = 80,6$; $SD = 12,8$, VG 4: $M = 75,0$; $SD = 7,5$). Die SUS-Werte pro Versuchsgruppe sind normalverteilt (Shapiro-Wilk-Test). Mittels ANOVA konnten keine signifikanten Unterschiede zwischen den Versuchsgruppen festgestellt werden.

Die qualitative Einschätzung der Versuchspersonen bezog sich häufig auf den VR- und CAVE-basierten Versuchsaufbau. Diesem schrieben die Probanden einen Videospielecharakter zu, der ihr Verhalten und insbesondere auch die Entscheidungsfindung beeinflusst haben könnte. In zukünftigen CAVE-basierten Studien sollte diesem Aspekt besondere Aufmerksamkeit geschenkt werden, indem möglichen Einflüssen durch realitätsnähere Gestaltung entgegengewirkt und / oder mögliche Einflüsse systematisch erfasst werden. Bezogen auf den Aufenthalt wurde die Anzahl von fünf Items als zu hoch und daher realitätsfern eingeschätzt. Hier muss ein besserer Kompromiss zwischen erfasster Datenmenge und realisti-

schem Nutzerverhalten gefunden werden. Einer Probandin missfiel, dass nach einigen Besuchen im Einzelhandel nur noch gastronomische Angebote empfohlen wurden. Dies ist ein deutlicher Hinweis darauf, dass die temporäre Score-Abwertung ähnlicher Items, stark von der Item-Domäne abhängt. Im gastronomischen Bereich kann von einem gestillten Hungergefühl ausgegangen werden, nachdem ein zugehöriges Item besucht wurde. In anderen Domänen ist der Besuch vieler ähnlicher Items möglicherweise gewünscht.

6 Diskussion und Fazit

Trotz verschiedener erkannter Verbesserungsmöglichkeiten ließ die Studie einen ersten positiven Eindruck des Assistenzsystems zur Aufenthaltsgestaltung erkennen, der sich in den Bewertungen der Gebrauchstauglichkeit widerspiegelt.

Die Ergebnisse bezüglich des Empfehlungssystems sind teils überraschend. Der Verzicht auf Tf-IDF-Gewichtung und Serendipitäts-Mechanismus ist numerisch signifikant genauer als der Einsatz beider Mechanismen einzeln. Beim Serendipitäts-Mechanismus entspricht dies der Erwartung, durch Tf-IDF-Gewichtung wurde allerdings eine Verbesserung der numerischen Genauigkeit erwartet. Analog ist auch die Tatsache, dass die Kombination beider Mechanismen numerisch signifikant genauer ist als der Einsatz nur eines der zusätzlichen Mechanismen bemerkenswert. Die Verringerung der numerischen Genauigkeit durch den Serendipitäts-Mechanismus könnte zwar durch die Gewichtung kompensiert worden sein, die alleinige Verwendung der Tf-IDF-Gewichtung sollte die numerische Genauigkeit allerdings steigern. Wie diese Ergebnisse vom Versuchsaufbau und der geringen Stichprobengröße beeinflusst wurden, ist in nachfolgenden Nutzerstudien zu klären.

VG 1 wurde bzgl. der numerischen und wahrgenommenen Genauigkeit am besten bewertet und erzielte zudem den höchsten SUS-Wert. Unerwartet ist, dass VG 2 bzgl. der Gebrauchstauglichkeit am schlechtesten bewertet wurde, obwohl die Empfehlungen numerisch und nach dem ResQue-Fragebogen am zweitbesten abschnitten. Ähnliches zeigt sich in den Ergebnissen der ResQue-Fragengruppe Nutzungsbereitschaft, die für VG 4 nur mit 1 bewertet wurde, während alle anderen VG ein Ergebnis von 4 erreichten. Diese Ergebnisse stehen im Widerspruch zur Annahme, dass die Gebrauchstauglichkeit des Empfehlungssystems positiv mit der Qualität seiner Empfehlungen korreliert. In weiteren Studien sollte dieser Zusammenhang weiter untersucht werden.

Ausgehend von den qualitativen Bewertungen des Versuchsaufbaus, kann ein Einfluss des VR-Versuchsaufbaus auf die Entscheidungsfindung der Probanden angenommen werden. Dieser Einfluss muss in künftigen Versuchen weiter eruiert werden, um die Reliabilität der Daten bzgl. der Entscheidungsfindung in VR-Umgebungen besser einschätzen zu können.

7 Literaturverzeichnis

- Adomavicius, G., Mobasher, B., Ricci, F. & Tuzhilin, A. (2011). Context-Aware Recommender Systems. *AI Magazine*, 32(3), 67.
- Bangor, A., Kortum, P., Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3), 114-123.
- Brooke, J. (1996). SUS: a 'quick and dirty' usability scale. In: Jordan, P. W. (Hrsg.): Usability evaluation in industry. London: Taylor & Francis, 189-194.
- Burke, R. (2007). Hybrid Web Recommender Systems. In Brusilovsky, P., Kobsa, A. & Nejdl, W. (Hrsg.): The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg: Springer, 377-408.
- Debernardis, S., Fiorentino, M., Gattullo, M., Monno, G. & Uva, A. E. (2014). Text Readability in Head-Worn Displays: Color and Style Optimization in Video versus Optical See-Through Devices. *IEEE Transactions on Visualization and Computer Graphics*, 20(1), 125-139.
- Gunawardana, A. & Shani, G. (2015). Evaluating Recommender Systems. In: Ricci, F., Rokach, L. & Shapira, B. (Hrsg.): Recommender Systems Handbook. New York: Springer, 265-308.
- Hüllermeier, E., Fürnkranz, J., Cheng, W. & Brinker, K. (2008). Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16), 1897-1916.
- Karrer, K., Glaser, C., Clemens, C. & Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG. *Der Mensch im Mittelpunkt technischer Systeme*, 8, 194-199.
- Klahold, A. (2009). Empfehlungssysteme: Grundlagen, Konzepte und Systeme. Wiesbaden: Vieweg + Teubner.
- Konstan, J. A. & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2), 101-123.
- Kotkov, D., Wang, S. & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111, 180-192.
- Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4), 2065-2073.
- Musto, C. (2010). Enhanced vector space models for content-based recommender systems. In Amatriain, X., Torrens, M., Resnick, P., Zanker, M., Guy, I. & Jaimés, A. (Hrsg.): Proceedings of the fourth ACM conference on Recommender systems. New York: ACM, 361-364.
- Oppermann, R. (1994). Introduction. In Oppermann, R. (Hrsg.): Adaptive user support: ergonomic design of manually and automatically adaptable software. Hillsdale: Lawrence Erlbaum Associates, 3-12.

- Pu, P., Chen, L. & Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4), 317-355.
- Ricci, F., Rokach, L. & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B. (Hrsg.): *Recommender Systems Handbook*. Boston: Springer, 1-35.
- Schafer, J. B., Frankowski, D., Herlocker, J. & Sen, S. (2007). Collaborative Filtering Recommender Systems. In Brusilovsky, P., Kobsa, A. & Nejdl, W. (Hrsg.): *The Adaptive Web: Methods and Strategies of Web Personalization*. Berlin, Heidelberg: Springer, 291-324.
- Zhao, Q., Adomavicius, G., Harper, F. M., Willemsen, M. & Konstan, J. A. (2017). Toward Better Interactions in Recommender Systems: Cycling and Serpentine Approaches for Top-N Item Lists. In Poltrock, S. & Lee, C. P. (Hrsg.): *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. New York: ACM, 1444-1453.