Mental Models, Explanations, Visualizations: Promoting User-Centered Qualities in Recommender Systems

Von der Fakultät für Ingenieurwissenschaften,

Abteilung Informatik und Angewandte Kognitionswissenschaft

der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte kumulative Dissertation

von

Johannes Kunkel
aus
Gladbeck

1. Gutachter: Prof. Dr.-Ing. Jürgen Ziegler
2. Gutachterin: Prof. Dr. Irene-Angelica Chounta

Tag der mündlichen Prüfung: 31. Oktober 2022

DOI: 10.17185/duepublico/78167
URN: urn:nbn:de:hbz:465-20230413-084211-6

Acknowledgments

This work has been influenced and supported directly or indirectly by so many wonderful people that it would exceed the space available here. Nevertheless, in the next lines I will try my best to express my humble gratitude to the most influential supervisors, researchers, and personal contacts, without whom this dissertation would not have been possible.

First and foremost, I would like to thank my PhD supervisor Jürgen Ziegler for his support. Your critical thinking, fatherly encouragement, and faith in my abilities helped me to rise above myself several times and in this way you greatly improved the quality of my conducted research. At the same time, you gave me the freedom to pursue my personal scientific interests. There have never been prohibitions on thinking in our fruitful discussions.

I would also like to take this opportunity to thank all other researchers who have accompanied me on my scientific journey. Gratitude goes especially to the supervisors of my Bachelor and Master thesis, Tim Hussein and Benedikt Loepp. Tim, you sparked my interest in recommender systems and helped me develop a foundation of scientific work routines that I could rely on in all the later years. Benedikt, you took over this baton and continued to guide me on my academic path. In many discussions as a co-author and on trips to conferences, you showed me what it means to be a devoted researcher. As my supervisor, you always engaged me on eye level—a quality of yours that has never failed to impress and encourage me.

My further thanks go to my co-PhD students who were in the same boat with me. Especially, to Thao Ngo, with whom I dived into the topic of mental models and who showed me a yet undiscovered perspective on my research area. I simply loved sitting with you on the floor of your office sorting drawings of kraken monsters. Aside from that, there were a tremendous number of other fine research colleagues, co-authors, reviewers, and professional contacts too numerous to mention individually here. Together we have laughed, argued, celebrated, and sometimes even cried. I thank you for every single experience we shared!

Finally, but by no means least, I would like to thank my family and friends for their support, care, and warm words over the past years. Of these, special thanks go to Lisa Götz for her compassion, patience, and unconditional love. Lisa, I cannot stress enough how much you helped me reach the goal of completing this thesis! Thank you so much. I would also like to thank my parents and especially my mother Katrin Kunkel. It is you who sparked my curiosity about the world and my drive to know how things work. You also taught me the attitude not to settle for unsatisfactory circumstances, but to take action and work to change them.

Thank you, all of you!

Annotation of the Papers Included in the Cumulus

Research on mental models:

- Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 183–191. ACM, New York, NY, USA, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394841
- Johannes Kunkel, Thao Ngo, Jürgen Ziegler, and Nicole Krämer. Identifying group-specific mental models of recommender systems: A novel quantitative approach. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Proceedings of the IFIP Conference on Human-Computer Interaction*, INTERACT 2021, pages 383–404. Springer International Publishing, 2021. ISBN 978-3-030-85610-6. doi: 10.1007/978-3-030-85610-6_23

Research on explanations:

- Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12. ACM, 2019. doi: 10.1145/3290605.3300717

Research on visualizations:

- Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 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, pages 3–15. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3025171.3025189
- Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. Newsviz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 126–135. ACM, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394869
- Johannes Kunkel and Jürgen Ziegler. A comparative study of item space visualizations for recommender systems. *International Journal of Human-Computer Studies*, 172, 2022. doi: https://doi.org/10.1016/j.ijhcs.2022.102987

Abstract

Recommender systems (RSs) are powerful tools that proactively suggest a set of personalized items to users. In doing so, they aim to predict the preferences of their users, wherein they are considered to be very accurate. In addition to algorithmic precision, user-centered qualities have recently been increasingly taken into account when evaluating the success of RSs. Examples for such qualities include the transparency of an RS, the control users are able to exert over their recommendations, and the means of exploring the item space in context of recommendations. However, research on aspects focused on human-computer interaction in RSs is still at a rather early stage. The main focus of the present thesis is to study and design RSs more holistically. In this regard, the *mental models* that users create of RSs are explored, explanations and their impact on user-centered variables of RSs are investigated, and techniques from *information visualization* (InfoVis) are applied to let users scrutinize the global context of their recommendations. The results of this research and the contributions I make to the state of the art in this context are described in greater detail below.

A key contribution of this thesis consists of the results of two studies that shed light on the mental models that users of RSs develop and how these models influence the users' perception of different system qualities. A key finding of the first, qualitative study is that many mental models tend to follow a procedural structure that can be used, for instance, as a template for designing explanations to promote transparency in RSs. In the second study, which relied on a larger sample and thus allowed quantitative conclusions, this type of procedurally structured mental models was found to correlate with a high perception of system transparency and confidence in the users' own comprehension of the inner workings of the system. Apart from that, some users seemed to *humanize* the RS, assigning attributes such as "social", "organic", and "empathic". Such a comprehension of the system was accompanied by higher levels of trust—a finding that may be leveraged by system designers. In general, mental models that deviate greatly from the actual functioning of the system should be corrected so that they do not lead to false expectations on the part of the users and hence to a potentially rejection of recommendations.

A prominent method for improving system transparency and thus the soundness of users' mental models is to provide textual explanations along with the recommendations. These explanations usually follow a very simple scheme based on similarity—especially in productive environments. To investigate implications of such simple explanations, another experiment contained in this thesis asked users to explain recommendations in their own words and compared them to explanations automatically generated by a system. The results indicate many benefits of providing more extensive explanations for recommendations, such as increased trust and higher perceived quality of recommendations. Another finding is that many participants, as opposed to the system, provided a broader context of the decision behind their recommendation.

The extent to which textual explanations can provide context for recommendations is limited, though. While a *local* context is relatively easy to explain textually—e.g. by linking recommendations to a user's preferences—it is difficult, if not impossible, to provide users with a *global* context. Such a global context would need to explain the relationship of recommendations to all other items in the dataset from which a RS selects its candidates. Comprehending such an *item space* at a global scale can unlock several beneficial properties of an RS, such as preventing filter bubbles, fostering creativity, and encouraging a user's self-development. In this thesis, I argue that to provide such a global context, RSs should go beyond explaining recommendations textually and better exploit the capabilities of computer systems compared to humans.

Three of the six papers included in this cumulative dissertation explore how methods of InfoVis can be applied to RSs to provide users with a global context of recommendations and how this affects the users' perception of these systems. One result of these studies is that even simple means of representing the item space can already successfully convey a sense of overview over the item space and provide transparency for recommendations. However, another finding is that artificial maps that distribute all items on a two-dimensional plane according to their similarity are a promising visualization style that can be used to deeply integrate means of interactively controlling recommendations into the visualization of the item space. Such maps have also been found to trigger user excitement, which can also influence the perception of recommendations. In another experiment, we found that a *treemap* can also be used as a control panel for a RSs. The results of this experiment further underline that treemaps can effectively alert their users to potential biases or blind spots in their preference profile. In this thesis, I discuss such implications of the InfoVis method to depict the item space of RSs.

Finally, in this thesis I take an elevated perspective on the findings of the papers contained and argue that researchers should consider user-centered aspects of RSs more holistically, for instance, in terms of the deep interconnectedness of perceptual variables. In this sense, I observed that the *user experience* of an application can influence as how novel recommendations are perceived to be, and that the degree of overview of the item space users are able to obtain can positively affect the perceived quality of recommendations. This thesis represents thus a further argument for looking at RSs from a highly user-centered viewpoint.

Keywords: Recommender Systems, Information Visualization, Mental Models, Explainable AI, Interactive Systems, Human-Computer Interaction

Kurzfassung

Empfehlungssysteme (ES) sind leistungsstarke Instrumente, die den Nutzern proaktiv eine Reihe von personalisierten Artikeln vorschlagen. Dabei zielen sie darauf ab, die Präferenzen ihrer Nutzer vorherzusagen, worin sie als sehr präzise angesehen werden. Neben der algorithmischen Akkuratesse werden in letzter Zeit auch zunehmend nutzerzentrierte Qualitäten bei der Bewertung des Erfolgs von ES berücksichtigt. Beispiele für solche Qualitäten sind die Transparenz eines RS, die Kontrolle, die Nutzer über ihre Empfehlungen ausüben können, und die Möglichkeit, den Item Space im Kontext von Empfehlungen zu erkunden. Die Forschung zu Aspekten, die sich auf die Mensch-Computer-Interaktion in ES konzentrieren, befindet sich jedoch noch in einem recht frühen Stadium. Das Hauptargument der vorliegenden Arbeit liegt in der ganzheitlicheren Untersuchung und Entwicklung von ES. In diesem Zusammenhang werden die *mentalen Modelle*, die Nutzer von RS erstellen, exploriert, Erklärungen und ihre Auswirkungen auf nutzerzentrierte Variablen von RS untersucht und Techniken der Informationsvisualisierung (InfoVis) angewandt, damit Nutzer den globalen Kontext ihrer Empfehlungen erkunden können. Die Ergebnisse dieser Forschung und die Beiträge, die ich in diesem Zusammenhang zum Stand der Wissenschaft leiste, werden im Folgenden ausführlicher beschrieben.

Der erste Beitrag dieser Arbeit besteht aus den Ergebnissen zweier Studien, die Aufschluss darüber geben, welche mentalen Modelle die Nutzer von ES entwickeln und wie diese Modelle die Wahrnehmung verschiedener Systemqualitäten durch die Nutzer beeinflussen. Ein zentrales Ergebnis der ersten, qualitativen Studie ist, dass viele mentale Modelle einer eher prozeduralen Struktur zu folgen, was z.B. als Vorlage für die Gestaltung von Erklärungen zur Förderung der Transparenz in ES verwendet werden kann. In der zweiten Studie, die sich auf eine größere Stichprobe stützte und somit quantitative Schlussfolgerungen ermöglichte, wurde festgestellt, dass diese Art von prozedural strukturierten mentalen Modellen mit einer hohen Wahrnehmung der Systemtransparenz und dem Vertrauen in das eigene Verständnis der inneren Funktionsweise des Systems korreliert. Darüber hinaus schienen einige Nutzer das ES zu *vermenschlichen*, indem sie ihm Attribute wie “sozial”, “organisch” und “empathisch” zuschrieben. Ein solches Verständnis des Systems ging mit einem höheren Maß an Vertrauen einher – eine Erkenntnis, die von Systemdesignern genutzt werden kann. Generell sollten mentale Modelle, die stark von der tatsächlichen Funktionsweise des Systems abweichen, korrigiert werden, damit sie nicht in falschen Erwartungen auf Seiten der Nutzer und somit einer möglichen Ablehnung von Empfehlungen resultieren.

Eine beliebte Methode zur Verbesserung der Systemtransparenz und damit der Korrektheit der mentalen Modelle der Benutzer ist die Bereitstellung von textuellen Erklärungen zusammen mit den Empfehlungen. Diese Erklärungen folgen in der Regel einem sehr einfachen Schema, das auf Ähnlichkeit beruht – insbesondere in produktiven Umgebungen. Um die Auswirkungen solcher einfachen Erklärungen zu untersuchen, wurden in einem weiteren Experiment im Rahmen dieser Arbeit Benutzer gebeten, Empfehlungen in ihren eigenen Worten zu erklären, und diese mit von einem System automatisch generierten Erklärungen verglichen. Die Ergebnisse deuten auf viele Vorteile hin, die sich aus ausführlicheren Erklärungen für Empfehlungen ergeben, wie z.B. einem höheren Vertrauen und einer höheren wahrgenommenen Qualität der Empfehlungen. Ein weiteres Ergebnis ist, dass viele Teilnehmer, im Gegensatz zum System, einen breiteren Kontext der Entscheidung hinter ihrer Empfehlung lieferten.

Der Umfang, in dem textuelle Erklärungen den Kontext für Empfehlungen liefern können, ist jedoch begrenzt. Während ein *lokaler* Kontext relativ leicht textuell erklärt werden kann – z.B. durch die Verknüpfung von Empfehlungen mit den Präferenzen eines Nutzers – ist es schwierig, wenn nicht unmöglich, den Nutzern einen *globalen* Kontext zu vermitteln. Ein solcher globaler Kontext müsste die Beziehung der Empfehlungen zu allen anderen Elementen in dem Datensatz erklären, aus dem ein ES seine Kandidaten auswählt. Das Verstehen eines solchen *Produktraums* auf globaler Ebene kann mehrere vorteilhafte Eigenschaften eines ES freischalten, wie z.B. die Vermeidung von Filterblasen, die Förderung von Kreativität und die Ermutigung zur Selbstentwicklung eines Nutzers. In dieser Arbeit argumentiere ich, dass ES, um einen solchen globalen Kontext zu bieten, über textuelle Erklärungen von Empfehlungen hinausgehen und die Fähigkeiten von Computersystemen im Vergleich zu Menschen besser ausnutzen sollten.

Drei der sechs in dieser kumulativen Dissertation enthaltenen Arbeiten untersuchen, wie Methoden der InfoVis auf ES angewandt werden können, um Nutzern einen globalen Kontext von Empfehlungen zu vermitteln und wie sich dies auf die nutzerzentrierten Qualitäten dieser Systeme auswirkt. Ein Ergebnis dieser Untersuchungen ist, dass bereits einfache Darstellungen des Produktraums erfolgreich einen Überblick über den Datenraum der Produkte vermitteln und Transparenz für Empfehlungen schaffen können. Ein weiteres Ergebnis ist jedoch auch, dass künstliche Karten, die alle Artikel auf einer zweidimensionalen Ebene entsprechend ihrer Ähnlichkeit verteilen, ein vielversprechender Visualisierungsstil sind, mit dem sich Mittel zur interaktiven Steuerung von Empfehlungen tief in die Visualisierung des Produktraums integrieren lassen. Es hat sich auch gezeigt, dass solche Karten bei den Nutzern Begeisterung auslösen können, was ebenfalls die Wahrnehmung von Empfehlungen beeinflussen kann. In einem weiteren Experiment haben wir herausgefunden, dass eine *Treemap* auch als Bedienfeld für ein ES verwendet werden kann. Die Ergebnisse dieses Experiments unterstreichen, dass Treemaps ihre Nutzer effektiv auf mögliche Verzerrungen oder blinde Flecken in ihrem Präferenzprofil hinweisen können. In dieser Arbeit diskutiere ich solche Implikationen der InfoVis-Methode zur Darstellung des Produktraums von ES.

Zusammenfassend nehme ich in dieser Arbeit eine erhöhtere Perspektive auf die vorgestellten Ergebnisse ein und argumentiere, dass Forscher nutzerzentrierte Aspekte von ES ganzheitlicher betrachten sollten, zum Beispiel im Hinblick auf die tiefe Verflechung von Wahrnehmungsvariablen. In diesem Sinne habe ich festgestellt, dass die *user experience* einer Anwendung einen Einfluss darauf haben kann, als wie *neuartig* Empfehlungen wahrgenommen werden, und dass der Grad des *Überblicks* über den Produkttraum, den Benutzer erhalten, die wahrgenommene *Qualität* der Empfehlungen beeinflussen kann. Die vorliegende Arbeit repräsentiert somit ein weiteres Argument für die Betrachtung von ES aus einer stark nutzerzentrierten Sicht.

Stichworte: Empfehlungssysteme, Informationsvisualisierung, mentale Modelle, erklärbare KI, interaktive Systeme, Mensch-Computer-Interaktion

Contents

Acronyms

CF *collaborative filtering*

DR *dimensionality reduction*

FC *feature congestion*

InfoVis *information visualization*

MAE *mean absolute error*

MDS *multidimensional scaling*

MF *matrix factorization*

NDCG *normalized discounted cumulative gain*

RMSE *root mean squared error*

RQ *research question*

RS *recommender system*

UX *user experience*

1 Introduction

Modern webshops, music streaming portals, travel booking sites and many other platforms leverage highly specialized intelligent systems to present their content tailored to the individual preferences of their customers. Since their introduction in the early 1990s, such *recommender systems* (RSs) have evolved into powerful instruments that aim at predicting a user's preferences and continue to be of great importance in science and industry [54, 138, 141]. Among other benefits, RSs have been shown to have the potential of increasing purchases, customer loyalty, and user satisfaction [140].

While the community of researchers and developers of RSs has long focused primarily on improving the accuracy of predicting a user's preferences, this is always a retrospective accuracy that may be outdated, misguided, or over-fitted and thus fails to take into account the user's actual current needs [74, 116]. As a result, many user-centered aspects of RSs have recently received some attention as well [115, 152, 157]. For example, it has been demonstrated that introducing more transparency into RSs is appreciated by users [153] and that can improve their trust in these systems and thus the acceptance of recommendations [68]. One of the most popular approaches aimed at making RSs more transparent is to provide textual explanations [68, 159, 177]. Another user-centered aspect of RSs is the degree of control users are able to exercise over the recommendations they receive [62, 65, 114]. For both qualities, it is pivotal to investigate what internal representation users create of the RS they interact with in order to prevent typical pitfalls of human-computer interaction.

Whenever humans interact with a digital system, they create a cognitive representation of that system. This *mental model* serves as functional simulation and helps predict the system's reactions and hence plan interaction steps to achieve the user's goal [128, 143]. An effective mental model is to a certain degree congruent with the actual system and a necessary prerequisite for any effective human-computer interaction. Initial studies have investigated which mental models users of intelligent systems develop [50, 51, 88]. It has been found that in addition to the effectiveness of human-computer interaction, the soundness of the mental models that users develop also influences their satisfaction with the system [37, 94, 121].

To help users build a more sound mental model and thus address aspects such as the degree of trust they place in an RS, recommendations are often presented along with textual explanations for them [68, 159, 177]. In the automated generation of explanations, RSs are constrained by the algorithm they use to calculate recommendations [21, 46]. Since many recommendation algorithms internally rely on similarities between items or users [91], a very common explanation style is to indicate a similarity between recommendations and a user's preferences or the item currently viewed (e.g. in the popular explanation style "*Users who bought... also bought...*" from Amazon). The effectiveness of such rather simple explanations has been questioned, though [21, 46].

Textual explanations usually explain the relationship of a recommendation to single items of the underlying item space[1]. Those items typically belong to the user's preference or, as in the aforementioned example of Amazon, the item being currently inspected. While providing a broader context would certainly be technically feasible in most cases—e.g. by indicating similarities of recommendations to a larger set or even the entire space of items—it remains questionable whether it would be practical to present users with such a large amount of information in the form of a text.

Yet, several downsides are known for situations where users are not aware of how their recommendations relate to the underlying item space. One of these downsides is that users may become trapped in *filter bubbles* [131] or *echo chambers* [42]. While it is still debated whether RSs are the primary cause for such situations or whether users voluntarily enter them [4], one way to counter them is to remind users what fraction of the entire item space they consume relative to the rest. Users who do not know how their recommendations relate to the rest of the item space may face other issues apart from filter bubbles. They may have a lower confidence in their choice [71], or be fearful of missing out [70]. They may also be unable to explore the item space more comprehensively and develop new preferences [87]. Finally, issues of fairness and ethics arise when users are nudged into a certain region of the item space [14, 28].

One way to alleviate these issues is to exploit the capability of digital systems to visualize large information domains. The discipline of *information visualization* (InfoVis) is a research field that is dedicated to, inter alia, the challenge of helping users make sense of large item spaces [22, 66, 151]. By leveraging InfoVis techniques, RSs users can be presented with the entire item space on a global scale, and be provided with means to explore it [e.g. 3, 44, 108]. In *TVLand* [44], for example, the domain of television shows is displayed as an artificial map in which regions of high and low preferences are highlighted in form of a *heat map*. Others have applied this map-based method to the domain of music [3] and university courses [108].

1.1 Research Questions and Contributions

From the current state of research outlined above, I derived four *research question*s (RQs) that form the basis for the work presented in this thesis. The RQs can be divided into two parts: Part I (RQ1 and RQ2) is concerned with studying how users perceive RSs and how this leads them to associate different attributes with these systems. Part II (RQ3 and RQ4) is concerned with how designers and engineers of RSs can foster this comprehension and leverage it to unlock desirable user-centered attributes of RSs. In particular, my four RQs are:

RQ1: Which mental models do users develop of an RS?

Problem Mental models are crucial for users to effectively use a digital system [50, 51, 88]. In the rare cases where mental models of RSs have been studied, they are typically described by the level of their soundness [50, 94] or individual aspects they contain [39, 51]. Capturing

[1] Here I define *item space* as "the set of all physical or digital objects that serve as recommendation candidates in a given database". Examples of item spaces are all movies available on *Netflix*, all tracks to choose from on *Spotify*, and all books physically available in a local library.

and describing a general mental model of an RS that is shared by multiple users has not yet been pursued, though. Insights in form of such an overarching model would provide clues regarding common presumptions that users have developed about the system and thus what directions should be taken to make RSs more transparent and usable. Moreover, the diversity of mental models in larger samples and across different RSs has not yet been studied. For example, it remains unclear whether users develop a separate model for each RS, or whether mental models are shared across systems.

Approach and Results To answer this RQ, two studies were conducted. In the first [127], we chose a qualitative approach and conducted interviews with Netflix users. Applying the *grounded theory* methodology [27, 155], we revealed a basic procedural mental model that all participants followed to some degree. This model comprises four steps: data acquisition, inference of user profiles, comparison of user profiles or items, and generation of recommendations. Apart from this, the mental models of participants were very diverse. This diversity was confirmed in a second study we conducted [103]. This second study was designed as an online survey and aimed to elicit mental models of a larger sample using card sorting. The resulting highly diverse card sorts yielded different distinct groups of participants who developed a similar mental model. We found three such groups: one with a *procedural* mental model, one with a *concept-based* mental model, and one group that developed a mental model comprising many *social* aspects.

Contributions The results of this research contribute to the respective field by providing in-depth insights into users' assumptions about how RSs work. For example, the four steps of the procedural mental model identified in the first study can be used to *anchor* explanations of RSs in users' existing comprehension. In addition, our second study demonstrates how mental models of RSs can be captured quantitatively in a broad sample. Another contribution lies in the three user groups we revealed in our results and that group users with a similar mental model. These groups can be leveraged to make RSs more transparent in the future, e.g. by determining to which of these groups the active user belongs and subsequently tailoring the displayed explanatory components to that group's model. A simpler approach would be to adhere to a procedural style of explaining recommendations, as we found that the largest group of participants in our experiment followed a procedural understanding of RSs, and it can thus be assumed that many users approaching an RS would intuitively understand such explanations.

RQ2: What is the relationship between mental models and users' perception of RSs?

Problem In addition to investigating which mental models users of RSs develop, a crucial question is *how* these models affect users' perception of RSs—especially in terms of whether connections exist between the mental model and user-centered aspects such as recommendation transparency, control, and trust. Some initial empirical results suggest that flawed mental models can lead to confusion [121] and that improving the soundness of mental models positively affects interaction efficiency and user satisfaction [94]. In tandem with the results of RQ1, the potential number of such correlations increases: not only the soundness of mental models could influence how the system is perceived and how one interacts with it, but also the specific content of a mental model, could lead to different attitudes towards an RS.

Approach and Results Methodologically, we addressed this RQ by administering questionnaires on user-centered aspects such as social presence, trusting beliefs, transparency, control, and perceived recommendation quality as part of the online survey we conducted to elicit users' mental models [103]. In contrast to typical interview-based approaches in mental model research, choosing this method allowed us to analyze the results quantitatively. Specifically, we tested for statistical differences between groups, which we identified through hierarchical clustering based on participants' card sorts. As a result, we found that *procedural* mental models were associated with a high perception of recommendation transparency and confidence in using the RS. However, participants who followed a more *concept-based* mental model did not perceive the RS to be very transparent. Finally, participants who focused on *social* aspects experienced the highest trust in the system, which we note is consistent with prior findings [25, 99].

Contributions The contributions of this research relate to designing systems that promote a type of mental model that can lead to more positive perceptions of an RS. In domains where, for instance, trust in the system is critical, social components should be emphasized. In certain situations, this might even involve altering the procedure for calculating recommendations. In this scenario, human actors could be involved in the process of generating recommendations, and this could then be communicated to users. All things considered, the results provide a basis for a stronger emphasis on human-in-the-loop in RSs.

RQ3: How do typical system-generated explanations compare to explanations based on natural language created by humans?

Problem One common way to increase the transparency of an RS and thus to improve a user's mental model is to provide textual explanations. The automated generation of such explanations depends on how the recommendations are calculated algorithmically. Since one of the most common ways of generating recommendations is based on item (or user) similarity, automatically generated explanations also often adhere to a similarity-based style [68, 159]. While richer approaches of explaining recommendations also exist [21, 35], there has been insufficient research on how conventional, similarity-based explanations generated by a system compare to explanations that humans would use to justify their recommendations. It can be assumed that richer explanations—especially those that mimic social interactions—improve trust due to the ability to form bonds with the RS [25, 99].

Approach and Results To answer this RQ, we designed and conducted an experiment in which two groups of users received movie recommendations [101]. In one group, users received recommendations from an RS combined with explanations based on item similarity (i.e. indicating the positively rated item most similar to the recommended item in each case). In the other user group, users were asked to recommend movies to each other and to briefly explain their choice. As a result, we found that while the system-generated recommendations were more accurate in terms of matching the target users preferences, humans as recommenders were better at explaining their choices. Interestingly, this superior explanatory ability completely evened out the difference in recommendation accuracy. We also found that the social presence emanating from natural language explanations was an influential mediator of the trust participants placed in the recommendation source (i.e. human or RS).

Contributions The results indicate that typical system-generated similarity-based explanations may be too shallow to be effective and that RSs could benefit from generating more complex explanations. Aside from perceived explanation quality, such improved explanations could potentially help system designers to increase other aspects such as a system's trustworthiness and even the experienced accuracy of recommendations. Thus, a more theoretical contribution is that attributes of RSs which are typically thought to be isolated from each other are more intertwined than is assumed.

RQ4: How can the underlying item space of an RS be visualized to users? How does the visualization style influence the users' perception of an RS?

Problem It has been shown several times that RSs can benefit from presenting users not only with recommendations, but also with a suitable representation of the entire item space [2, 108, 118]. To achieve this, methods from InfoVis have often been used. Thereby many of the visualizations used are based on a map metaphor, because maps are able to display large item spaces while remaining intuitively comprehensible [45, 120]. Yet, the interaction with recommendations, e.g. in the form of supporting efficient control of preferences, is often not directly supported by such visualizations. It is also underexplored how different visualization types compare to each other and what benefits, for example, rather complex map visualizations bear compared to simpler and more common representations of the item space of an RS.

Approach and Results To answer these two RQs, we implemented six different interfaces for visualizing the item space of RSs. One of these interfaces used a list, one used tiles with sliders, two used treemaps, and the last two were based on a map metaphor. Of these six interfaces, we compared one of the treemap interfaces with the tile-based slider interface (*NewsViz* [102]), one of the map interface with the other treemap interface and the list (*MusicExplorationApp* [97]), and conducted a usability study with the remaining map interface (*MovieLandscape* [98]). Especially with *NewsViz* and *MovieLandscape*, we demonstrate how to deeply integrate the functionality of an RS and the way users can control recommendations into the visualization of the underlying data. When compared to a baseline, we found that users experienced a higher degree of control. In the *MusicExplorationApp*, we showed that the visualization style influenced overall user experience, which was highest for the interface based on a map metaphor.

Contributions One of the primary implications of our experiments with different item space visualizations for RSs is that InfoVis methods, when properly integrated, can increase the control users are able to exert over recommendations. On the other hand, the results also indicate that even simple means of exploring the item space can be sufficient and provide high transparency of recommendations. However, the user experience seems to benefit from more complex map visualizations, which in turn can influence other perceptional aspects of RSs such as the perceived novelty of recommendations.

1.2 Outline

This dissertation is organized as follows: The next chapter (Chapter 2) briefly summarizes the relevant literature and discusses background not contained in the papers. Chapter 3 then presents the research contained in the cumulus. In Chapter 4, I reflect on the findings of my research and bring them in relation to the ongoing scientific discourse in the respective areas. Finally, Chapter 5 provides a brief summary of my major contributions and suggests a few next steps to advance the research presented here.

2 Background and Related Work

Since their advent in the 1990s, RSs have become widely used systems that proactively suggest travel destinations, music tracks, commercial products, and many other items to users [54, 138, 141]. By providing individually tailored sets of recommendations to users, RSs can reduce the cognitive burden of manually coping with information overload, while from the recommendation provider's perspective, RSs can help increase purchases, user satisfaction, and customer loyalty [140].

There are several ways to generate recommendations. Most prominent among them are *content-based* algorithms and algorithms based on *collaborative filtering* (CF). As the name suggests, content-based recommendation algorithms focus the content of items to determine a set of relevant recommendations for the active user [123]. In case the content of items is available and can be analyzed appropriately, a content-based RS can operate directly on the items themselves, e.g. by using word embedding techniques such as *word2vec* [117] on text documents such as news articles. On the other hand, in case the content of items cannot be analyzed easily, metadata can be used to operate a content-based RS. This can be achieved, for instance, by adding information from an external data source such as *Linked Open Data*, as done by Passant [133]. To recommend suitable items to a user, they need to be associated with a user profile that contains the same attributes as those known for the items. Then, a relevance score can be calculated for the items, and the ones with the highest scores can be selected as recommendations [123]. CF, on the other hand, does not depend on the availability of content information for items, but is performed on interaction data between items and users [91]. These interaction data can consist of explicit ratings that users consciously provide (e.g. elicited on a 5-star rating scale), or of implicit ratings that are obtained from other interactions (e.g. click-through rates or dwell times on a web page). Based on these ratings, similarities between items or users can be calculated. One strategy for identifying potential items to recommend to the active user is to select those items that similar users have rated highly in the past [91].

One widely known approach for CF is *matrix factorization* (MF) [90, 158]. In MF, the user-item matrix $\mathbf{R} \in \mathbb{R}^{|U| \times |I|}$, which contains interaction data for all users U and items I, is decomposed into $\mathbf{P} \in \mathbb{R}^{|U| \times f}$ and $\mathbf{Q} \in \mathbb{R}^{|I| \times f}$, two low-rank matrices with f representing a given number of factors where typically $f \lll |U|$ and $f \lll |I|$. There are several approaches for this decomposition, for instance, *alternating least squares* and *singular value decomposition* [90]. The predicted rating $\hat{r}_{ui}$ for user $u \in U$ and item $i \in I$ is calculated by the inner product of the user's factor vector $\vec{p_u} \in \mathbb{R}^f$ and the item's factor vector $\vec{q_i} \in \mathbb{R}^f$, stored in $\mathbf{P}_u$ and $\mathbf{Q}_i$, respectively. The factors in $\mathbf{P}$ and $\mathbf{Q}$ are referred to as *latent* because they are assumed to contain hidden semantics that are distilled from the underlying ratings [34, 100, 125, 142].

In order to evaluate the performance of an RS, a very popular method is to conduct offline experiments [5, 58, 158]. This type of evaluation is often the easiest to perform and entails relatively low costs, as it does not require an extensive user study design or subject acquisition.

To estimate the performance of a recommendation algorithm offline, various techniques can be used. One of the most popular metrics is the *mean absolute error* (MAE), which quantifies the error in the rating predictions of an RS. Based on an existing dataset of interaction data, a small portion of users and items is retained as a test set $\mathcal{T}$. Then, the recommendation algorithm is performed on the remaining data of $\mathbf{R}$ and rating predictions for all user-item pairs $(u, i) \in \mathcal{T}$ are computed, for instance, by MF as described above. Then the MAE is defined by $\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} |r_{ui} - \hat{r}_{ui}|$. A variant of the MAE is to compute the *root mean squared error* (RMSE), where larger errors are disproportionately penalized. There are several other approaches to measure the accuracy of recommendation algorithms; for example, some consider the ranking of recommendations instead of its numeric prediction accuracy, such as the *normalized discounted cumulative gain* (NDCG) [72].

Regardless of the specific measurement approach, offline experiments are particularly useful for ensuring the computational precision of an algorithm during its development, but also for determining a pre-selection of suitable existing recommendation algorithms for an RS. On the other hand, offline experiments are insufficient to be used as the only instrument for evaluating the overall performance of an RS [58, 115, 157]. One of the downsides of offline experiments is that they rely solely on historical data, and thus on the assumption that the used dataset models the users' behavior after the system is deployed well enough. Another reason why offline experiments are inappropriate for measuring the overall performance of an RS is that there are several other aspects of an RS that affect its success besides its algorithmic precision. Many of these aspects focus on the user and how they perceive and interact with the RS.

2.1 User-Centered Quality Aspects of Recommender Systems

Apart from the ability of an RS to make as accurate predictions as possible about a user's future ratings and thus about which items are appropriate recommendation candidates, many other factors have been identified that influence the success of these systems [85, 89, 115, 152, 157]. Some of these qualities, *beyond accuracy*, pertain to the composition of the set of recommendations returned by an RS. When a user is presented with very accurate recommendations that all fit well with their tastes, they may face *choice overload* [10, 70]. In an online survey, Bollen et al. found that the users' choice satisfaction consists of an interplay between the variables of *attractiveness* (i.e. recommendations that match the user's taste), *variety* (i.e. recommendations that consist of very different items), and *choice difficulty.* This indicates that large sets of recommendations containing only items that are very attractive to the user may not be the most satisfying recommendations for them. To tackle choice overload effects and other drawbacks of very homogeneous recommendation sets, there are several approaches on how to achieve a higher diversity by an RS [18, 76, 169]. These approaches range from re-ranking an existing recommendation list to approaches that rely, for instance, on clustering the user's taste profile and generating recommendations for each cluster separately [18]. Closely related to the diversity of recommendations is their *novelty.* Both aspects can also be found in common frameworks for evaluating RSs [e.g. 86, 136].

While recommendation diversity and novelty can be addressed algorithmically and even measured offline to some extent, improving other user-centered aspects in RSs requires a careful design of how recommendations are displayed to users and how they can interact with the system. One of these aspects is the degree of *control* users are able to exercise over their recommendations. Over the years, interactive RSs have evolved into their own branch of research [15, 65, 74]. All interactions with an RS can be basically organized into two categories: *preference elicitation* and *interaction during result presentation* [74]. Traditionally, preferences are elicited by RSs in form of implicit or explicit ratings on a single item base [73, 83, 154]. Yet, also more efficient solutions exist where users can, for instance, make a binary choice for a set of items [105] or rate item clusters represented by tags [20]. However, these methods are typically performed *before* recommendations are made (e.g. at cold start), which can be difficult for users, for instance, because they lack the domain knowledge to explicitly articulate their preferences [112].

Critiquing in the context of RSs [112, 171, 173], as a technique that is performed during presentation of recommendations, takes a different route. In this approach, users are asked to indicate directions for an RS to take *after* they have seen and/or consumed a recommended item. In the seminal work of Vig et al. [173], the authors present *Movie Tuner*, an RS that allows users to critique movie recommendations based on tags, letting them, for instance, indicate that they want recommendations that are *less violent* than a current movie recommendation. In addition, many interactive systems have been proposed in which preference elicitation, presentation of recommendations, and general interactive components are deeply interwoven into a sophisticated system [e.g. 2, 55, 106, 118]. Improving interactive control over an RS, when implemented appropriately, has been shown to increase user satisfaction [38, 62, 74], perceptions of recommendation accuracy [130], and users' trust in the system [38]. When an RS provides appropriate means for its users to exercise control, this can also promote curiosity and willingness to explore an item domain beyond the areas, a user is already familiar with [165, 166]. In any case, control in the field of RSs is closely related to the comprehensibility of these systems [61, 65, 87].

Increased comprehensibility of RSs can be achieved by improving the recommendation *transparency* [15, 68, 153]. Providing recommendations that users can understand can be done in different ways. In the *Movie Tuner* application, mentioned above, recommendations are explained based on tags that the recommended item is assigned with and that the user has rated positively in the past [172]. A link to the user's preferences is also exploited in the *TasteWeights* system [12]. In *TasteWeights*, recommendations and user preferences are connected through an intermediate layer that explains how both are related. This could be, for instance, that they belong to a similar genre or were liked by a friend on Facebook. One goal of the application *TalkExplorer* [170] is to make the source of recommendations more visible to users. In *TalkExplorer*, recommendations are visualized as nodes in the active user's neighborhood. The resulting graph helps to identify the influence on recommendations in hybrid RSs settings. Making hybrid recommendations more transparent is also the goal of *Relevance Tuner+* by Tsai and Brusilovsky [166]. In this application, the influence of each of five recommendation sources is displayed as a relevance score, which can also be adjusted interactively. In addition to these examples, several other applications have been introduced that aim at making recommendations more transparent. Among them are approaches that use rather exotic techniques, such as displaying a user's preferences in the form of comic-like avatars [9]. Transparency, when implemented appropriately, has been shown to improve the acceptance of recommendations [52, 68], their perceived quality [101], and the trust users place in the RS [159, 176].

Knijnenburg et al. [87] first proposed the notion of *self-actualization* in the context of RSs. In contrast to the user-centered properties of RSs discussed above, self-actualization lies on a higher level of abstraction, as systems that follow this paradigm “support users in developing, exploring, and understanding their own unique tastes and preferences” [87]. As such, these systems aim to capture user preferences more holistically, account for temporary preferences, and promote personal growth. Consequently, transparency and control over recommendations are integral components of RSs for self-actualization [87, 156]. One advantage of RSs for self-actualization is that they support human creativity and are an effective countermeasure to *filter bubbles*, a controversially discussed situation in which users are “trapped” in their personalized taste profiles, which constantly reinforce themselves without users being able to leave them [131, 139]. A main concern in the context of filter bubbles is that users are only exposed to very homogeneous content, which can lead to very extreme opinions and polarization of society [28]. One way to alleviate this is to present more heterogeneous content, for example by generating more diverse recommendation sets. This can be achieved algorithmically [18, 76] or by providing interactive tools to make recommendations more heterogeneous [53, 77]. In this context, it has been shown that users who are more aware of how their own preference profile within an RS relates to the entire item space are less likely to have blind-spots and therefore to become trapped in filter bubbles [95].

One of the goals of RSs for self-actualization is to “support rather than replace decision-making” [87]. This supportive role of RSs can also be found in the notion of *supertools* recently introduced by Shneiderman [152]. In his work, he summarizes the ongoing trend of *human-centered AI* and its potential to support “human self-efficacy, creativity, responsibility, and social connections”. In doing so, he further emphasizes the need for modern intelligent systems to consider human users more holistically than has been the case in the past. In human-centered AI, as Shneiderman understands it, the intelligent system is neither a fully system-controlled *intelligent agent* nor a fully user-controlled *tool*. Instead, it is a *supertool* that provides a high degree of control to its users, but leverages all the available capabilities of modern AI to amplify their abilities. For Shneiderman, appropriately designed RSs are examples of such supertools because they provide decision support to their users, but do not take those decisions away from them. As such, they have to be *comprehensible*, *predictable*, and *controllable*, which is consistent with the increasing attention to user-centered aspects in RSs research and design discussed above.

2.2 Explanation Methods of Recommendations

One way to help users understand recommendations and increase their transparency is to provide textual explanations. Explanations can help bridge the information gap between RS and users, thus helping them to judge the quality of recommendations and decide whether they are worth to be followed [160]. When implemented appropriately, explanations for recommendations have been shown to improve overall user satisfaction [46], the trust users put in an RS [68, 159], and the perceived transparency [6]. On the other hand, however, Berkovsky et al. found that explanations based on the similarity between preferred and recommended items failed to convey trust [6]. Bilgic and Mooney [7] even found that explanations can have negative effects. More precisely, Bilgic and Mooney observed that users misjudged the quality of recommended items when given similarity-based explanations.

As a suggestion for structuring the discussion about explanations in RSs, Ain et al. [1] provide a conceptualization framework. This framework contains five dimensions, including the target audience of the explanation (e.g. are the explanation receivers individual users or groups?), the goal that the explanation interface pursues (e.g. to improve trust in the RS or to educate the user about the inner workings of the system), and the style of explanation used. A popular explanation style for recommendations is to reveal their relationship to the user's previously expressed preferences [68, 159] or to highlight content features that the user is likely to enjoy [7]. Recently, more complex types of textual explanations have emerged. For example, these can be based on natural language [21] or on user-created product reviews [35]. Tintarev and Masthoff [160] propose to organize the different explanation styles in three levels. The first level explains recommendations based on a single user. Explanations at this level have in common that they are very close to the raw data stored for the active user. An example of an explanation style in this category is the style used at Netflix, where recommendations are related to a movie that the user has rated positively in the past. The second level of explanations *contextualizes* the data of the active user. Explanations based on CF most often fall into this category because they mention the active user's context or *neighborhood*. A very popular example of this style follows the structure "*Users who bought... also bought...*", which is used by Amazon, for example. Finally, as third level, Tintarev and Masthoff list explanations to supports users in their self-actualization. That is, these explanations can be used to pursue a number of higher-level goals, such as "*discover the unexplored*" [156], which is closely related to *serendipity* [76, 110] and describes the ability of users to develop new preferences, as Tintarev and Masthoff explain.

2.3 Mental Models of Recommender Systems

In cognitive science, mental models, as internal *representations* of the external world, are a pivotal concept as they dictate how we perceive and interact with this external, *represented world* [128, 143]. Mental models have been studied since the early 1980s and describe "the way people understand some domain of knowledge" such as how liquids behave [49]. They are constructed through interaction with the external world and are thus subjective in nature, incomplete, uncertain, and possibly flawed [128]. Since some cognitive effort is required to develop a mental model, users will subconsciously reuse once constructed models whenever possible. This reuse of mental models, while cognitively efficient, can lead to misaligned mental models, for example, when the knowledge domain in which a model is used only appears to be similar to the domain in which it was created [128, 129]. As cognitive representation of an external domain of knowledge, mental models of digital systems have also been studied. As such, they are closely related to *folk theories*, which have been used to describe the imperfect theories users create about digital systems [30, 39, 126].

Regardless of whether they are referred to as folk theories or mental models, the internal representations of digital systems need to match the actual system functioning to some degree in order to be useful for predicting its behavior and thus for using the system effectively. While the imperfection of mental models is a cognitive fact, models that are too misaligned can cause errors in human-computer interaction. The *gap* between the users' mental model and the actual system behavior is closely related to the *gulfs* of *evaluation* and *execution*—two well-known concepts in human-computer interaction [129]. When the mental model fails in correctly predicting the current state of a system and thus in correctly interpreting the system output, a large *gulf of evaluation* is present. This situation leads to confusion and impedes users from achieving their goals with the system. Similarly, users cannot achieve their goals when a large *gulf of execution* is present. In this situation, the mental model does not contain the correct information to tell the user what actions to take in order to achieve their goal. In other words, the *gulf of evaluation* impedes the user from understanding *why* a system behaves the way it does, while the *gulf of execution* impedes the user from understanding *how* to effectively control the system.

More recently, mental models have received some attention in the research community of intelligent systems as well. Tullio et al., for instance, observed that users possessed some basic comprehension of conceptual and technical aspects of machine learning, such as decision trees and pattern recognition, when being confronted with an intelligent systems about which they had no prior knowledge [168]. Makri et al. [109] observed that users constructed a shared mental model for their experience with different types of document search interfaces. As a result, users' mental models were flawed and meagerly developed, which in turn resulted in the aforementioned *gulfs*. Muramatsu and Pratt [121] observed that flaws in the mental models of intelligent systems can cause confusion in the interpretation of search engine results. In a study on correcting misaligned mental models, Kulesza et al. [94] showed that improving the quality of users' mental models can increase the effectiveness of using a music RS. In line with this, Makri et al. suggested that intelligent systems should be developed that support users in creating better mental models [109]. Makri et al. further recommend that the exploration of data should be more actively supported. Eiband et al. [37] also focus on the development of intelligent systems that support users in creating better mental models. More precisely, they present a stage-based development process that takes into account the mental models that users develop in each development iteration. Finally, Kodama et al. suggest to start improving mental models early in people's contact with intelligent systems and to teach the theoretical and practical concepts behind search engines in schools [88].

In order to study mental models empirically and to capture the knowledge of a group of users about a certain subject, qualitative methods have usually been applied. In his seminal work entitled "*Some Observations on Mental Models*", Norman [128] presents results of an interview study in which users of calculators were asked to *think aloud* while performing different simple tasks with the calculator and verbalize their mental model afterwards. In a similar manner, Kodama et al. [88] elicited the mental models of search engines in a sample of middle school students. Instead of an interaction task during which they should think aloud, Kodama et al. asked the students to draw how they think the search engine works internally and then to verbalize what they drew in brief retrospective interviews. Apart from these exploratory qualitative studies, a few experiments have been conducted that approach mental models using quantitative methods. Thereby these experiments often focus on the *effects* of the mental models users hold on different dependent perceptional variables. Kulesza et al. [94], for instance, measured the soundness of participants' mental models using multiple-choice questions. Others employed conceptual techniques such as card sorting in order to identify the mental models of larger samples than is possible through face-to-face interviews [26, 104]. In contrast to small sized qualitative studies, investigating mental models in larger samples can help reveal the diversity of mental models that co-exist in a given group of users [51].

2.4 Visualizations in Recommender Systems

It has often been argued that the typical linear presentation of recommendations as a ranked list is not optimal in several regards [23, 71, 135]. In industry, carousel interfaces are also frequently used (e.g. at Netflix or Amazon), which are essentially a two-dimensional version of a ranked list. This makes the optimization problem of the underlying RS more complex, since not only the horizontal ranking within a recommendation list has to be considered, but also the vertical ranking of different lists of recommendations [33, 40]. Also, the evaluation of carousels needs to be approached slightly differently than individual recommendation lists, since their vertical position among the other recommendation lists needs to be considered as well [41]. In comparison to ranked lists carousels have been found to result in fewer interaction steps before a user finds a desired item [137].

From a user-centered perspective, the use of recommendation presentation techniques that go beyond linear ranked lists can unlock several desirable properties of RSs. For example, the perceived transparency and acceptance of recommendations can be increased by using bar charts [36]. This is supported by findings of Chen and Tsoi [23], who compared presenting recommendations in the form of a conventional vertically displayed list, a two-dimensional grid, and a circularly organized pie interface. As a result, they observed that the presentation of recommendations as pie and grid were preferred by users compared to the ranked list. Chen and Tsoi also found that these visualization styles resulted in a more even distribution of attention across the recommendations, and that the intention to follow the recommendations was increased. In regard of the aforementioned concept of self-actualization, Guesmi et al. [56] followed a human-centered design process to develop prototypes that visualize user models of intelligent systems to promote different user-centered goals such as exploration, comprehension, and socializing. Suggestions of their design process include bubble charts and Venn diagrams. Parra et al. [132] compared a list-based condition with their system *SetFusion*, which uses a Venn diagram to display the influence of different algorithms in a hybrid RS setting. They showed that *SetFusion* was perceived as more engaging. Similarly, Kouki et al. [92] observed an increased user experience when they used Venn diagrams to display the influence of different sources in a hybrid RS. In contrast, Tsai and Brusilovsky [166] used a Venn diagram in the aforementioned system *Relevance Tuner+* to explain similarity between users as an overlap of word clouds. Besides results on other forms of visualization, Tsai and Brusilovsky observed that the word cloud was experienced as enjoyable by users. Moreover, the word clouds helped users to better understand the recommendations. In a previous experiment, we found that even subtle visual cues about the source of a recommendation can foster the user's trust in an RS [99].

Multiple times, techniques from the research discipline of InfoVis [16, 22, 82] have been applied to RSs. Katarya et al. [80], for instance, use a *treemap* [150] to display the set of recommendations to users. A treemap is also used by Chang et al. [19], who use it to visualize a user's search queries. User studies with both applications revealed that users were satisfied with this style of visualization. Another InfoVis technique that has been applied to RSs are *node-link graphs*. For example, Gretarsson et al. [55] visualize a user's preference profile and its relationship to recommendations in form of a multilayer graph in which the user, their preferences, social peers, and recommendations are displayed as nodes.

In addition, users can manipulate the position of these nodes and thus influence their relevance for recommendations. Results from a user study conducted by Gretarsson et al. suggest that the application can improve the degree of control users are able to exercise over the RS. Gajdusek and Peska [43] also use a node-link graph, but to display clusters of a music dataset and the artists contained. While Gajdusek and Peska did not evaluate their application, Petridis et al. [135] conducted a user study with their prototype *TastePaths*, which is also set in the music domain and displays artists as nodes in a graph. Among other results, they found that *TastePaths* educates its users about the relationships between artists, thus helping them discover interesting music they were not aware of. Improving the users' mental model through this visualization also helped them to understand recommendations. In the *MusiCube* application [145], users can select musical features (e.g. tempo, acoustic texture) by adjusting the axes of a scatter plot of songs. In a small user study, participants indicated that *MusiCube* is an effective interface for receiving music recommendations. Tsai and Brusilovsky [167] introduce *Scatter Viz*, which uses a similar visualization. *Scatter Viz* is a social RS that recommends academic scholars based on a research profile. As with *MusiCube*, users of *Scatter Viz* can customize the features to be displayed as the axes of the scatterplot. In a user study, Tsai and Brusilovsky found that *Scatter Viz* performed better on several user-centered dimensions such as *trust*, *supportiveness*, and *intention to reuse*, compared to a baseline system that used a list to present recommendations. In contrast to these scatterplots, which have clear labels on their axes, other types of *maps* have been used to visualize the item space of an RS [2, 44, 84, 108]. Some of these approaches are discussed in greater detail further below.

Another reason for using more sophisticated means of presenting recommendations to users is to make them aware of the underlying item space from which an RS selects its candidates. This is important, for example, to prevent the aforementioned *filter bubbles* [131] or *echo chambers* [42]. As discussed above, one countermeasure against filter bubbles is to diversify recommendations and expose users not only to items that match their preferences known by an RS. While this may motivate users to explore areas of the item space with which they are unfamiliar, it may be difficult for them to anticipate the outcome of exploratory actions that take them beyond their typical user profile—potentially resulting in a wider gulf of execution [77]. Kangasrääsiö et al. [77] have demonstrated how to mitigate this by letting users iteratively explore the item space based on a spatial map metaphor. A similar solution is introduced by Tsai, who uses a two-dimensional scatterplot to display recommendations [164]. Tsai observed that when using this visualization, users apply a broader exploration pattern and thus discover more diverse items than with a traditional list-based presentation of recommendations [164]. Nagulendra and Vassileva [124] presented recommendations in their RS within a stylized bubble surrounded by topics currently not represented in the recommendations. As a result, Nagulendra and Vassileva found that users developed a good understanding of the filter bubble concept. With a similar objective, Tintarev et al. [161] displayed a user's blind-spots as part of a chord diagram. In a user study, they found that this chord diagram outperformed a simple bar chart in terms of conveying the blind-spots to users. However, in these examples, users are presented with a comparatively small number of items, and it remains unclear whether these visualization approaches would be able to display all of the items in the database of an RS. Yet, this would be beneficial in making users aware of all of their available options and thus of possible blind-spots.

Recently, Jannach and Adomavicius [71] listed "*help users explore or understand the item space*" as one goal of a *purposeful* RS. Apart from making users aware of potential blind-spots, providing a global overview of the item space has been shown to be beneficial in several ways. For example, when users do not understand the item space and thus the options available to them, they may have less confidence in their choice [71] or even regret it after making a selection [70]. It can also support users comprehend their own unique preferences, deepen existing preferences, and discover new ones, thus serving as another means to promote user self-actualization [135]. To achieve this, a key aspect of RSs for self-actualization is to enable and motivate their users to explore the item space around their existing tastes and in areas they have not yet made any experiences with.

To explore the item space beyond personalized recommendations, numerous interactive systems have been developed that depict recommendations in their global context [e.g. 2, 44, 108]. To achieve this, it is necessary to visualize large amounts of data effectively. Especially maps have proven useful to adequately convey the overall distribution of an item space. The application *nepTune* by Knees et al. [84], for instance, allows users to explore an item space of music as a three-dimensional geographic landscape. Knees et al. showed in a users study that participants enjoyed using *nepTune* to explore and discover music. Another map-based application in the music domain is presented by Andjelkovic et al. [3] and named *MoodPlay*. In contrast to *nepTune*, users in *MoodPlay* are presented with their listening history visualized as a path in the item space map. When comparing their application to a baseline without such visualization, Andjelkovic et al. found increased scores for recommendation transparency and degree of control. Ma et al. [108] similarly compared their map-based application *CourseQ* to a baseline and also found that their interface resulted in significantly higher values for transparency of recommendations. Over time, many other map visualizations for RSs have been presented, indicating that they are a valuable tool for users to explore large item spaces and the recommendations therein [24, 44, 75, 135, 148].

The popularity of a map-based form of visualizing large item spaces is likely due to its ease of understanding and intuitiveness in use. One aspect responsible for this intuitive comprehensibility of maps is that thinking spatially is innate to humans [45]. Without any additional explanations, spatial distances in a map are intuitively understood as semantic relatedness, with smaller distances between two markers representing a higher similarity—an observation that is also known as the "*first law of geography*" [162]. It has been found that the first law of geography also applies to artificial maps that display non-geographic data [120]. Probably for this reason, maps have proven to be superior in representing similarities between items than, for instance, treemaps [8]. Maps have also shown to be useful in providing users with an overview of the item space being depicted [163], and can thus be used well as entry point for users into an InfoVis system [151].

A central challenge in the design of item space maps is the visualization of the typically vast number of items stored in the database of an RS. To tackle this, a comprehensive abstraction of the underlying item space has to be found that supports users in orienting themselves without overwhelming them with information. This is typically achieved either by displaying representative sample items [84, 111, 146, 148, 174], labels with meta-information (e.g. music genres, or social tags) [3, 24, 81, 96], or a combination thereof [44].

3 Synopsis of the Research Papers Included in this Dissertation

This chapter briefly explains the conducted research that is contained in the six papers included in this dissertation. First, the two papers that deal with investigating mental models of RSs are presented [103, 127]. Then, I summarize the research conducted on human-generated explanations in comparison to computer-generated explanations [101]. Finally, the three papers about the use of InfoVis in RSs are presented [97, 98, 102]. In the subsequent Chapter 4, the results summarized here are discussed in terms of answering the RQs introduced in Section 1.1.

3.1 Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study

Problem While there is a considerable body of work that addresses user-centered qualities of RSs [e.g. 15, 136, 152], the cognitive representations that users develop of these systems have rarely been investigated. Uncovering such *mental models* is crucial for studying RSs from a user perspective, as the internal representation of an external system determines how users perceive and interact with it [128, 143]. As a consequence, mental models can be considered as a source of the qualities that user-centered research investigates in RSs. Revealing users' mental models enables the identification and thus elimination of false assumptions, potential misunderstandings, and other common pitfalls of human-computer interaction [39, 50, 126]. Studies conducted in this area have rarely investigated the specific content of mental models, though. This would be interesting, however, to study the interaction between human and RSs, for instance, in terms of which parts of mental models are flawed and should therefore be corrected.

Approach With the goal of uncovering the *mental models* that users of RSs create, we conducted a qualitative interview study [127]. This study followed the *grounded theory* methodology [27, 155]. Grounded theory is a technique for developing *theories* about a subject in a highly exploratory manner. This methodology is thus particularly suitable for research areas where not much preliminary work has been carried out.

To investigate which mental models users of RSs develop, we conducted semi-structured interviews with 10 regular Netflix users. We selected Netflix as the RS because it contains clearly visible recommendations (e.g. the "top picks for you" category) and is very popular. Therefore, we assumed that by the time of the interview, participants were likely to have already developed a mental model about this RS. As preparation, we developed an interview guide with open-ended questions such as "*Which part of Netflix do you think is personalized?*" and

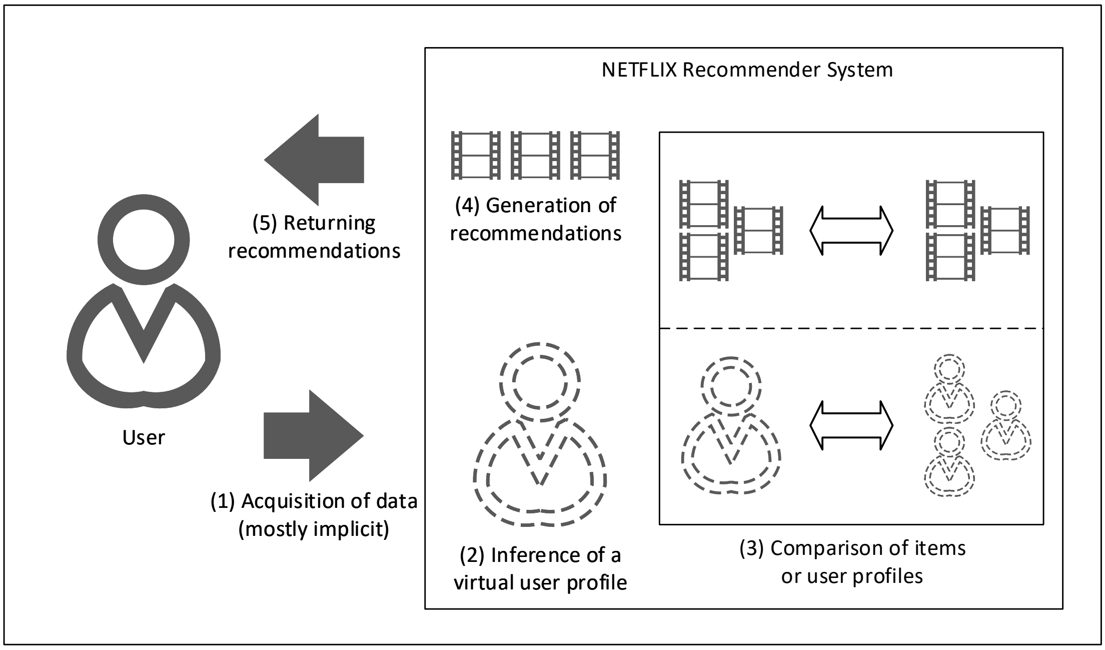

NETFLIX Recommender System
User
(5) Returning recommendations
(1) Acquisition of data (mostly implicit)
(4) Generation of recommendations
(2) Inference of a virtual user profile
(3) Comparison of items or user profiles

Figure 3.1: The central process, which we found to some extent in every mental model of our participants. Based on the collected data of a user (1), a virtual user profile is constructed (2). This profile is brought into relation with other profiles or items of the underlying dataset (3) and analyzed to generate recommendations (4), which are finally presented back to the user (5).

"*What data do you think are used for the personalization in Netflix?*". In addition to these questions, the interview also included a brief session in which participants were asked to *think aloud* while logging into their Netflix account and searching for a movie they would want to watch. Subsequently, they were instructed to imagine that they did not like that movie and to verbalize how they would communicate this to the system. With this task, we wanted to encourage participants to reflect on how they could steer the RS into a certain direction. As final task, we asked participants to draw their mental model of the RS in Netflix, a task that has also been utilized previously in such studies [e.g. in 32, 88].

A central tool of grounded theory is *theoretical sampling*: A certain group of participants is interviewed until no further variation can be found in the participants' responses. Then, another group of participants is focused or the experiment is concluded. This method requires researchers to alternate between conducting interviews and analyzing the results. In our case, we conducted three sampling phases: 1) "typical" Netflix users, i.e. participants in the primary target audience of the streaming platform; 2) participants with comparatively high or low levels of technical knowledge; and 3) participants who use the explicit rating feature in Netflix (i.e. thumbs up/thumbs down) to deliberately steer their recommendations in a particular direction. After this third phase, we concluded our sampling as we found no further concepts in the data that had not already been mentioned—or in grounded theory terminology: as *theoretical saturation* had been reached.

Results Analysis of the interviews revealed several main *concepts* in the mental models that were present in the utterances of multiple participants. These concepts were: *centrality of self*, *item-based vs. user-based recommending*, and a *technical vs. metaphorical* style of the mental model. *Centrality of self* was a concept that was very evident in some participants' models: their mental model was entirely organized around themselves.

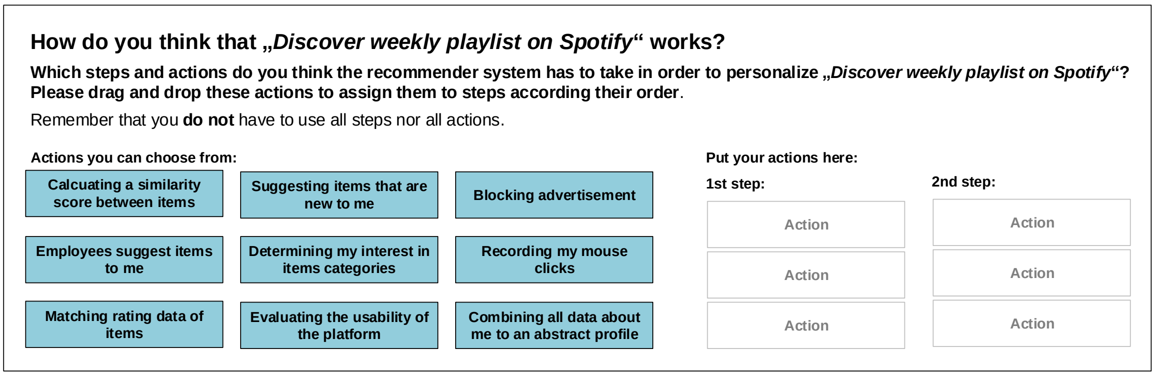
These participants also often assumed that the RS operated in a *user-based* manner, for example, by inferring similarities in the users' watch histories. Others, however, did not see themselves or the user in general as so central and focused more on the relationship between items, which we accounted for with the concept of *item-based recommending.* Mental models that followed this perspective were often more *technical* and focused on the algorithmic processes that an RS uses internally. Yet, other participants described their mental model in a more *metaphorical* language, using, for instance, the image of a tentacled monster providing recommendations with its many arms. Although participants varied widely in these concepts, we nevertheless found a central process model that all participants followed in their descriptions (see Figure 3.1).

From the above observations, we derive several implications for the development of RSs. For example, the basic mental model as presented in Figure 3.1 can serve as a guide for developers who want to explain recommendations to their users. Our research suggests that these steps are likely to be present in the mental model of all, or at least most, users. The aforementioned concept of *centrality of self* could also be used to explain recommendations. This implies that the relationship between the active user's preferences and the recommendations should be made explicit so that the user can understand how the recommendations are linked to their preferences. As a further implication, we suggest that erroneous mental models should be corrected. During the interviews, we received several responses expressing severe uncertainty and trust issues towards the RS, which we ascribe to a *mystification* of the underlying algorithms.

Contributions In summary, the contributions of our work to the ongoing discussion on user-centered qualities of RSs are twofold: 1) The exploration of mental models makes a theoretical contribution by providing deep insights into the way users view and interact with an RS. The general structure in Figure 3.1 is an example of these insights. 2) The research also contributes practically in that we derive several implications for future development of RSs.

3.2 Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach

Problem The research presented above provides some qualitative insights into the mental models of a relatively small sample of users of one particular RS. However, we were also interested in the prevalence and diversity of mental models *across* different RS and in a larger sample. In addition, we also wanted to analyze the relationships between mental models and users' attitudes towards these systems in a quantitative manner.



How do you think that „*Discover weekly playlist on Spotify*“ works?

Which steps and actions do you think the recommender system has to take in order to personalize „*Discover weekly playlist on Spotify*“? Please drag and drop these actions to assign them to steps according their order.

Remember that you **do not** have to use all steps nor all actions.

Actions you can choose from:
Calcuating a similarity score between items
Suggesting items that are new to me
Blocking advertisement
Employees suggest items to me
Determining my interest in items categories
Recording my mouse clicks
Matching rating data of items
Evaluating the usability of the platform
Combining all data about me to an abstract profile
Put your actions here:
1st step:
Action
Action
Action
2nd step:
Action
Action
Action

Figure 3.2: Excerpt from the card sorting task as presented to participants. The text about the particular RS (here: "Discover weekly playlist on Spotify") depended on which system the respective participant had chosen as reference system at the beginning of the survey.

Approach To achieve this, we designed and conducted a second experiment [103]. For this experiment, we developed a novel approach to elicit users' mental models of RSs using the *card sorting* technique in conjunction with a questionnaire. With card sorting, we chose a method that has previously been found to be suitable for eliciting mental models [26, 104]. We adapted this method to the functioning of RSs and extended it with questions about different attitudes and perceptual variables. This novel experimental setting enabled us to reach a wide range of participants and thus to make statements not only about a common, very abstract mental model as in the qualitative study described in Section 3.1, but also about the diversity of different mental models evident in a large sample.

For our study, we collected responses from $N = 170$ participants. Each of them was asked to chose one RS as a reference system. For this purpose, we provided a list of eight different systems to choose from. Examples are "Top pics for you on Netflix", "Video recommendations on YouTube", and "Discover weekly playlist on Spotify". The system chosen was displayed at several points in the questionnaire of the survey. In the card sorting task, participants were asked to select as many cards from a given set as they felt appropriate to express their understanding of how their respective RS works and sort them in up to seven consecutive steps (an excerpt from the presentation of this task is shown in Figure 3.2). Participants could choose from a total of 35 cards:

- 16 *action cards*: These cards represent typical paradigms used in RSs that we have gathered from previous mental model studies, from the RS literature, and from our own expertise. Examples of action cards are: "Recording my mouse clicks", "Analyzing content of items", and "Presenting items that other users liked in the past".
- 12 *distractor cards*: With these cards, we aimed at enabling users to add cards that deviate from the *ground truth* about how RSs work and thus to express misconceptions or faulty assumptions in their mental model. Examples for distractor cards are: "Employees suggest items for me" and "Evaluating the usability of the platform".
- 4 *question mark cards*: On these cards, a question mark was depicted to allow participants

to indicate that they believe something is happening at this step, but that they are too unsure to use a more specific card.

- 3 *open cards*: Finally, we added open cards on which participants could write their own text. In this way, they could add actions that we had not anticipated to be part of their mental model.

As user-centered qualities of the various RSs, we used questionnaire constructs for *social presence* [48], *trusting beliefs* [113], *transparency* [136], *control* [136], *perceived usefulness* [136], *recommendation quality* [86], and *perceived system effectiveness* [86]. In addition, there were some self-formulated questions about participants' confidence in the soundness of their mental model, whether it followed a *technical or metaphorical* style, and their *prior knowledge* about RSs.

Results For the analysis of the resulting card sorts, we compared all participants in terms of how they sorted their cards so that we obtained a matrix that contained similarities between each pair of participants. We then applied *Ward's method* [122, 175], an *agglomerative* (i.e. bottom-up) hierarchical clustering approach. In this way, we obtained three distinct groups of participants, organized according to their way of sorting the cards, and thus according to their mental models of RSs. Subsequently, we applied Ward's method again to cluster the cards *within* each of these groups so that we could also draw qualitative conclusions about the mental model of each group of participants. Together with the questionnaire results from each of the three groups, we identified the following three mental models:

Concept-based mental model ($N = 66$): Participants with a concept-based mental model arranged the cards thematically rather than chronologically. In this group, the aforementioned concepts of *item-based vs. user-based recommending* were particularly present. The clusters of cards in this group, for instance, indicate that one cluster adheres to *processing of the user model* and another to *processing of items*. Participants in this group reported having the least prior knowledge of how RSs work and that they perceived recommendations as not very transparent. Consistent with their concept-based mental model, their model was more technical than metaphorical.

Procedural mental model ($N = 79$): Unlike the first group, the second group of participants had a procedural mental model and arranged the cards in chronological order. This process begins with *data collection*, followed by *comparing items and users*. Subsequently, inferences are made about the user's interests, and finally, the recommendations are presented back to the user. We note that this group fairly directly mirrors the basic mental model we identified in the previous qualitative study (Section 3.1). Participants in this group followed a technical model and were highly confident that their mental model corresponded to how the system actually worked. Probably as a result, this group experienced the highest transparency of recommendations.

Social-focused mental model ($N=25$): The card sorts of this group did not adhere to any discernible procedural or concept-based pattern. Instead, their mental model was quite confused and seemed to follow a social and metaphorical style: As part of the questionnaire, participants assigned *metaphorical* attributes to the RS and indicated to experience the highest *social* *presence*. They also expressed the highest trust in the system—probably a result of assigning humanized attributes to the system.

Several implications for the design of RSs emerge from the results of the user study. As in the previous study, we found many uncertain and erroneous mental models, which is why we echo our suggestion made above to better explain RS to users. In particular, we emphasize that the transparency conveyed by such explanations should match the actual system functioning to some extent. This can prevent users from developing a false sense of understanding that is not based on facts and could thus lead to wrong assumptions and expectations. We stress that such a misguided form of understanding can lead to the aforementioned *gulfs* of human-computer interaction. Our results also reinforce that most users follow a procedural way of comprehending an RS, which we recommend as a template for explanations—especially in situations where nothing is known about the active user's mental model. Finally, our results indicate that a humanized mental model is related to increased levels of trust in the RS, and we hence suggest using a vivid, metaphorical language to instill trust in RS users.

Contributions With the work presented in this paper, we contribute to the research on how users comprehend RSs by presenting a novel card sorting setting that captures the entire processing chain of RSs. The result is a method for revealing mental models of broad samples that provides a foundation for future user studies. Based on the findings of our study, we propose several implications that make practical contributions to the future design of intelligent systems.

3.3 Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems

Problem Transparency and trust play a pivotal role in RSs and determine whether users are willing to follow recommendations or reject them. Textual explanations are a popular approach to increase transparency and trust in an RS [68, 159]. Automatically generated textual explanations range from rather simple similarity-based styles to richer explanations based on natural language [6, 21, 35]. However, the common similarity-based explanation styles have not yet been compared to human-formulated explanations for recommendations. There is evidence that, in particular, the trust a user places in an RS might benefit from explanations that follow a richer style [25, 99].

Approach We conducted an experiment showing how the ability to justify recommendations affects trust in the recommender and that even the perceived quality of recommendations depends on how well an RS is able to explain its recommendations [101]. More specifically, we compared rather simple explanations based on CF (i.e. "*Users who bought... also bought...*"), as are common for automated RSs, with more extensive explanations formulated by humans for other humans.

For our study, we developed a system that allowed participants to create and receive movie recommendations. In order to ensure that participants were able to consume each recommended movie, we restricted participation to those who had an Amazon Prime account. This left us

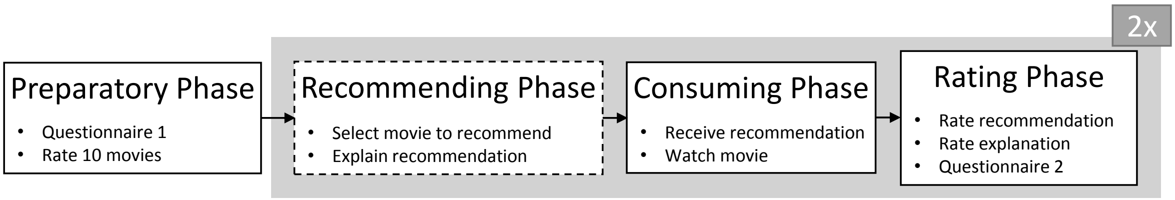


Figure 3.3: Phases of the study in which we compared recommendations and their explanations between users (personal condition) and an RS (impersonal condition). The *recommending phase* took place only in the personal condition. *Recommending phase*, *consuming phase*, and *rating phase* were performed twice during the course of the experiment.

with a total sample size of $N = 93$. As database, we relied on the widely used Movielens 20M ratings dataset[2], which we pre-filtered to those movies that were available for free on Amazon Prime at the time of the study. To compare human and automated sources of recommendations and explanations, each participant was randomly assigned to either a *personal* condition, in which they acted as both recommendation provider and recipient, or an *impersonal* condition, in which they received recommendations from an automated system. The procedure of our study was divided into four phases (Figure 3.3): *preparatory phase*, *recommending phase*, *consuming phase*, and *rating phase*.

In the *preparatory phase*, participants were asked to complete a questionnaire on general demographics and their attitudes toward technology, e.g. their *disposition to trust* [113] and their *general trust in technology* [86]. In this phase, they were also asked to rate 10 movies. The *recommending phase* was only performed by the participants in the personal condition. Here, they had to make a recommendation for another participant in this condition and explain their choice. To assess the other participant's taste in movies, they were presented with their previously rated movies. In the *consuming phase*, all participants watched the movie recommended to them and rated the recommendation and explanation in the *rating phase*. In this phase, participants were also asked to complete a second questionnaire about their trust in the recommendation source (i.e. their *trusting beliefs* and *trusting intentions* [113]) and the degree of perceived *social presence* [47]. The *recommending*, *consuming*, and *rating* phases were conducted twice. The entire course of our study spanned a total of two weeks.

Results In a direct comparison between conditions, we found that participants with other humans as recommenders perceived a higher *social presence* and experienced an increased *explanation quality* compared to participants in the impersonal condition. The automated RS, on the other hand, provided recommendations of higher quality. A mediation analysis revealed that this superior ability of the system to suggest items with high precision was entirely counterbalanced by its inferior ability to explain its suggestions. In other words, the recommendations of a human source are perceived to be similarly accurate to those of an RS only

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

because humans are superior in their ability to justify their choice. A *counterfactual analysis* helped us quantify this observation: if an RS would be able to explain its recommendations as well as humans, the quality of its recommendations would increase by about 13 % or 0.5 stars on a 5-star rating scale. Apart from this salient relationship between explanation and recommendation quality, we were able to uncover further relationships between dependent variables. For instance, we found that the variables *recommendation quality*, *explanation quality*, and *social presence* affected participants' trust in the source of recommendations.

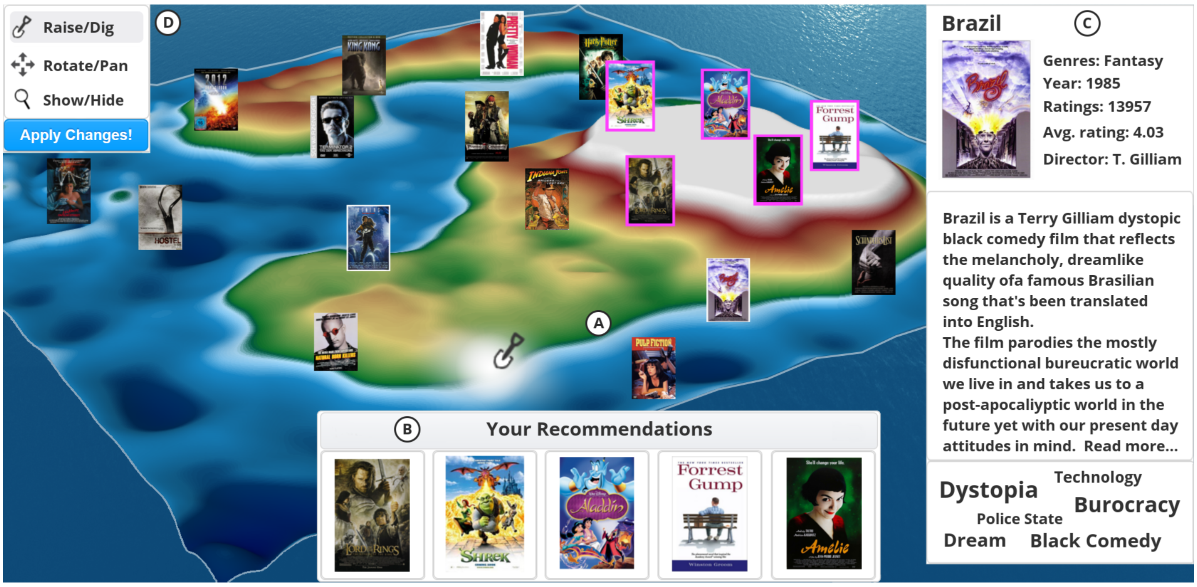
Although it was not part of the main focus of the data analysis, we also performed a cursory qualitative review of the collected explanations. In doing so, we found that the concept of similarity, on which the system-generated explanations were also based, was used fairly often in the explanations created by humans. However, participants of our experiment contextualized their explanations better, for instance, in mentioning multiple similar movies or in using a richer combination of different explanation styles (e.g. similarity-based and content-based).

Contributions In summary, the work presented in this paper contributes to the research on user-centered qualities of RSs in different ways: The conducted user study demonstrates that personal and impersonal recommendation sources differ in their ability to make recommendations and to justify their choices. This leads to further direct and indirect effects. For example, the perceived quality of recommendations depends on a recommender's ability to explain their choice, which, together with other user-centered variables such as perceived *social presence*, influences how much a user trusts the source of recommendations. We especially emphasize that typical simple explanations based on item similarity are too shallow to be fully effective.

3.4 A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering (MovieLandscape)

Problem The human-generated explanations in the study discussed in the previous section were probably superior than the system-generated explanations because they better *contextualized* the recommendations. In Section 2.4, I additionally argued that providing such context at a *global* scale may have several benefits such as prevention of filter bubbles and blind spots [131], a better choice confidence [71], and support for *self-actualization* [135]. At the same time, contemporary RSs that use CF take a quite opaque algorithmic approach—especially when they rely on model-based techniques such as MF [34, 90]. It is this technique of CF that is behind the simplistic explanations based on the neighborhood of a recommendation ("*Users who bought... also bought...*") that we criticized in the paper discussed in the previous section.

Approach To tackle the issues above, we propose a novel system (Figure 3.4) [98]. The pivotal component of this appliction, which we later named *MovieLandscape*, is a two-dimensional plane (A) on which about 10 000 movies are disseminated based on their similarity. In order to avoid overwhelming users with information, only representative movies for the different areas of the map are displayed. However, these hidden movies can also be explored using an interactive tool from the tool palette (D): after selecting the *Show/Hide* tool, users can click



Brazil

Genres: Fantasy
Year: 1985
Ratings: 13957
Avg. rating: 4.03
Director: T. Gilliam

Brazil is a Terry Gilliam dystopic black comedy film that reflects the melancholy, dreamlike quality ofa famous Brasilian song that's been translated into English.
The film parodies the mostly disfunctional bureucratic world we live in and takes us to a post-apocalyptic world in the future yet with our present day attitudes in mind. Read more...

Figure 3.4: Screenshot of the *MovieLandscape* application. The central item space visualization (A) shows representative movies samples for different areas of the map. The elevations represent the active user's preferences, which are located in the eastern part of the map. Recommendations that match these preferences are shown in a dedicated recommendation area (B), but also within the map with a magenta colored margin. The movie *Brazil* is currently selected and its details are displayed in the details area (C). Users can select tools from a palette (D) to interact with the visualization and rotate it, zoom in or out, show or hide sample items, and change the preference profile of the landscape, thus triggering a recalculation of recommendations.

on an empty space, whereupon the item located at that position will be added to the set of samples. On this map, the elevations indicate the active user's preference profile, so that areas of high preference are displayed as hills and areas of low preference are displayed as valleys (or in this case, as ocean depths). Another tool allows the user to reshape this landscape if they want to receive different recommendations. Recommendations are calculated after each interaction with the elevation profile and are highlighted on the map and displayed in the form of a list (B). Hovering the mouse cursor over a movie poster displays details about it in the details area (C).

The process of how *MovieLandscape* is technically realized can be seen in Figure 3.5. Based on the same dataset of about 20 million movie ratings as mentioned in Section 3.3, MF is performed as described in Chapter 2. This results in feature vectors for each item and user (1a). Subsequently, the dimensionality is reduced by applying *multidimensional scaling* (MDS) [11] to produce a two-dimensional space (1b). While this could already be presented to users, it would most likely overwhelm them due to information overload. Hence, we determined a small set of representative items as follows. First, we clustered the two-dimensional movie distribution using *k-means* (1c). For each of the 30 resulting clusters, we automatically selected a sample movie based on how representative it is for all the movies in the corresponding cluster (1d). We determined the score for representativeness analogously to the approach explained

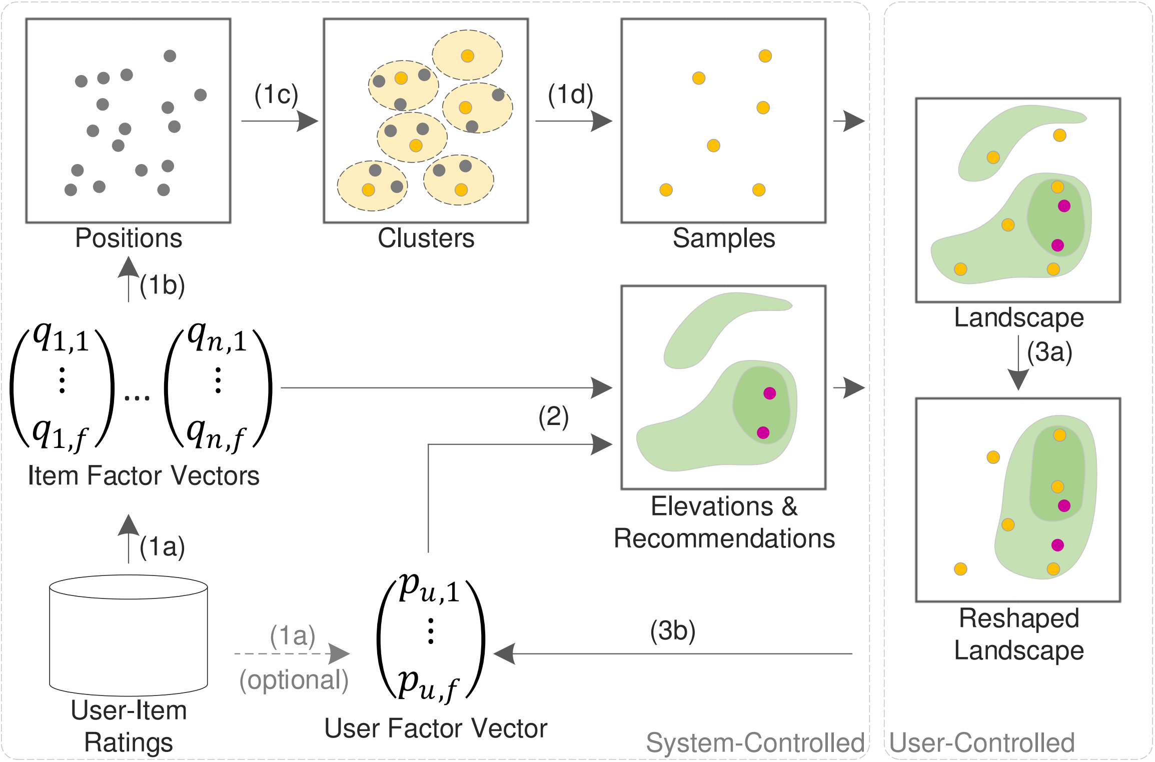


Figure 3.5: The central process of the *MovieLandscape* application: Latent factors for all users and items are determined using MF (1a). A two-dimensional distribution is generated from the factors for items using MDS (1b). Subsequently, k-means clustering is performed (1c) and a representative item is selected for each cluster (1d) to obtain a comprehensible item space representation for users. In case preferences of the active user are already known, they are displayed as hills and valleys (2). The user can now reshape the landscape according to their taste (3a), which results in a recalculation of their user vector and recommendations (3b).

by Loepp et al. [105].

Based on this two-dimensional item space map, the preference profile is presented to the users. For this purpose, the active user's rating predictions as calculated by MF are translated to elevations in the map: areas with high predicted preferences are depicted as hills while areas with low preferences are depicted as valleys (2). A virtual tool allows users to reshape this terrain according to their preferences (3a). Such functionality would be particularly useful when dealing with short-term preferences (e.g. when searching for a movie to watch in a group), at cold-start (i.e. when no information about the user is yet available), or when a user's preferences have changed over time. Any interactive reshaping of the landscape triggers an immediate recalculation of recommendations.

Results In a user study, we evaluated the *MovieLandscape* application with $N = 32$ participants. After a brief introduction to the functionality of the application, participants were asked to perform three consecutive tasks. In the first task, which also served as introduction, participants were asked to explore the item space and to find some pre-determined movies. In the second task, they were asked to create a novel preference profile from scratch, i.e. starting with a flat landscape. Finally, in the third task, we presented each participant with one of three pre-determined user profiles and asked them to change the landscape according to their personal needs. Besides others, we found that recommendations were perceived as transparent and controllable by participants. They also attested a high *user experience* (UX) to the system and indicated that they had fun while using it.

Contributions In conclusion, the work presented in this section contributes by demonstrating how RSs based on rather abstract model-based CF can be deeply integrated into an exploration tool that makes use of a map metaphor. This approach takes advantage of the algorithmic accuracy of MF while exploiting the often reported *hidden semantics* of such latent factor models [34, 100, 142]. As a result, users gain control over an RS by interacting directly with a three-dimensional representation of their user preference vector.

3.5 Depicting and Controlling Preference Profiles Using Interactive Treemaps in News Recommender Systems (NewsViz)

Problem RSs in the domain of news articles face some particular challenges. Even though their effects are sometimes discussed controversially [59, 119], *filter bubbles* [131] or *echo chambers* [42] have been associated with ideological segregation [42], populism [31], and the distribution of conspiracy theories [29]. One countermeasure for such intellectual impoverishment is to provide a broader overview of the underlying information domain to alert users to potential blind spots [60] and allow them to explore items beyond their usual preferences [87]. It has also been argued that providing citizens with a broad and unbiased viewpoint is a pivotal requirement for democracy [67].

Another aspect that is particularly challenging for RSs in the domain of news is that new articles appear frequently and that user preferences change often [78]. As a result, it is particularly important to provide users with control over their preference profile, which is also desired by users [61].

Approach With the *NewsViz* application [102], we address these challenges by providing an overview of the categories and news sources of an aggregated news feed while allowing users to control the personalization of this feed. In *NewsViz*, we utilize a treemap to visualize the news feed composition (Figure 3.6), which is an InfoVis technique that is particularly suited for displaying the proportion of attributes of tree cells to their root [150]. The treemap cells in *NewsViz* represent news categories (e.g. politics, sports, business) whose size can be configured by the user to control the proportional number of respective articles in the personalized news feed. Within each of these cells, news sources are presented as a second hierarchical treemap level. In this way, the user can configure how their personalized news feed is composed in terms of news publishers. The underlying recommendation algorithm follows the procedure known as *post-filtering strategy* [78]: The treemap composition regulates how the news feed

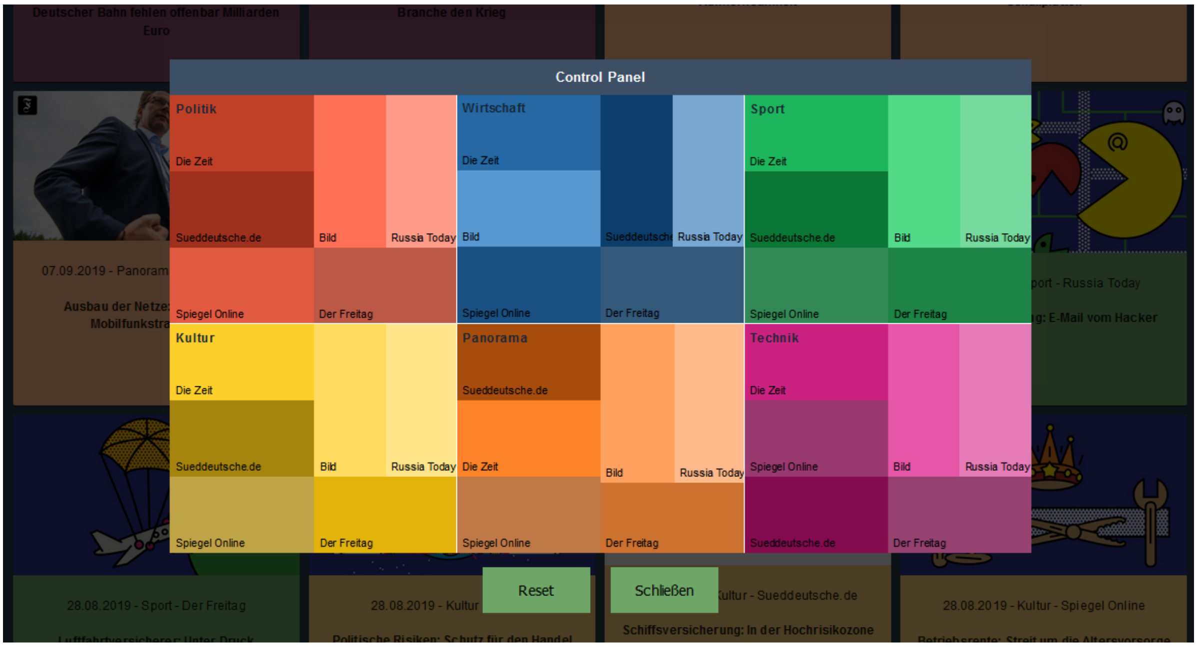


Figure 3.6: Screenshot of the *NewsViz* application. By hovering the cursor over any news category or source and moving the mouse wheel, the user can adjust the impact of the respective category or source on a personalized news feed (visible in the background). The cell sizes indicate the proportion of news articles in the feed that are associated with the corresponding news category or source.

is composed of news categories and news sources proportional to the corresponding treemap cells. Subsequent to this filtering process, the remaining feed is sorted by recency.

To test whether *NewsViz* is an appropriate tool for letting users control their personalized news feed and raising their awareness for possible biases in a user profile, we conducted a comparative user study ($N = 63$). To this end, we developed a second prototype that we used as a baseline to compare *NewsViz* with. This second prototype had exactly the same functionality as *NewsViz*, but used a slider-based interface to let users examine and control their preference profile. This allowed us to isolate effects that the treemap visualization had on user-centered qualities of the RS. More precisely, we elicited the variables *transparency* [136], *overall satisfaction* [136], *interaction adequacy* [136], which we used to measure partcipants' perceived degree of control over the RS, *effort* [86], *recommendation quality* [86], and *system effectiveness* [86]. To these we added some self-formulated questions to survey the degree of overview of the item space that participants were able to obtain. The experiment was composed of two tasks. In the first task, participants were asked to use the interface to configure the news feed according to their preferences. As the second task, participants were presented with screenshots of pre-configured profiles and asked to indicate the most influential news publisher for each of these screenshots.

Results As a result, we found a significant difference between conditions in the degree of control participants were able to exert over their news feed—i.e. recommendations. Participants who were presented with *NewsViz* perceived significantly higher levels of control over the composition of their news feed. Using *structural equation modeling*, we were furthermore able to reveal the large influence of the overview participants perceived. This variable positively influenced the transparency of recommendations, the quality of recommendations, and the system effectiveness. As a result of the second task, we found that participants were more frequently successful in determining the most influential source when using *NewsViz* than when using the slider-based variant. The same was true for the time it took participants to identify the most influential source: participants who were presented with *NewsViz* were significantly better and faster in determining the most influential source in their news feed.

Contributions The results of our experiment showed that the treemap helped users control the composition of their personalized news feed. Even though sliders are certainly the more common interactive widget, users of *NewsViz* quickly learned how to use the treemap and felt confident using it. The superiority of the treemap interface in providing control compared to the slider-based variant was most likely due to the better visualization of the proportional influence of news categories or sources. We conclude that representing the proportions as area sizes of the treemap cells was an intuitive way to convey the overall composition of the news feed. As such, like *MovieLandscape*, we consider *NewsViz* to be an example of a deep integration of control into the functionality of an RS. In the second task, more participants were able to identify the most influential news source by examining screenshots of the treemap compared to sliders. Consequently, we consider our proposed *NewsViz* interface to be an effective way to make users aware of potential biases and filter bubbles in media-aggregating websites.

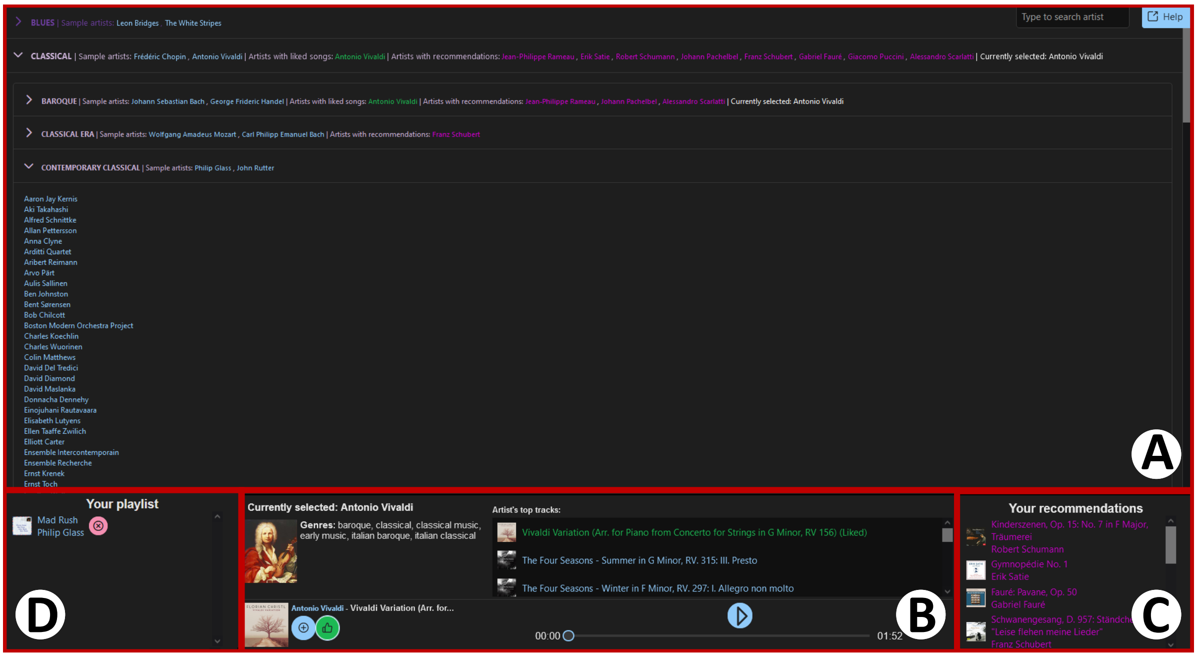
3.6 Investigating the Influence of Different Item Space Visualizations on Recommender Systems (MusicExplorationApp)

Problem Several applications have been presented that seek to leverage the ability of InfoVis to visualize large datasets in order to address user-centered challenges of RSs. The *MovieLandscape* and *NewsViz* applications represent two such approaches. To date, however, different InfoVis and non-InfoVis interfaces in the context of RSs have rarely been compared in user experiments.

Approach To address this gap, we developed a set of prototypes for the domain of music [97]. These prototypes depict a dataset of about 9 000 music artists and about 90 000 songs as *List*, *Treemap*, and *Map* (Figure 3.7, Figure 3.8, and Figure 3.9).

To create this dataset, we manually set up a hierarchy with two levels of music genres (13 top-level genres, each with 5 subgenres). Subsequently, we used the Spotify API[3] to retrieved artists and their most popular songs for each genre. To provide users with an orientation about the music in each genre, we determined two representative sample artists for each of them.

[3] `https://developer.spotify.com/documentation/web-api/`



BLUES | Sample artists: Leon Bridges , The White Stripes

CLASSICAL | Sample artists: Frédéric Chopin , Antonio Vivaldi | Artists with liked songs: Antonio Vivaldi | Artists with recommendations: Jean-Philippe Rameau , Erik Satie , Robert Schumann , Johann Pachelbel , Franz Schubert , Gabriel Fauré , Giacomo Puccini , Alessandro Scarlatti | Currently selected: Antonio Vivaldi

BAROQUE | Sample artists: Johann Sebastian Bach , George Frideric Handel | Artists with liked songs: Antonio Vivaldi | Artists with recommendations: Jean-Philippe Rameau , Johann Pachelbel , Alessandro Scarlatti | Currently selected: Antonio Vivaldi

CLASSICAL ERA | Sample artists: Wolfgang Amadeus Mozart , Carl Philipp Emanuel Bach | Artists with recommendations: Franz Schubert

CONTEMPORARY CLASSICAL | Sample artists: Philip Glass , John Rutter

Aaron Jay Kernis
Aki Takahashi
Alfred Schnittke
Allan Pettersson
Anna Clyne
Arditti Quartet
Aribert Reimann
Arvo Pärt
Aulis Sallinen
Ben Johnston
Bent Sørensen
Bob Chilcott
Boston Modern Orchestra Project
Charles Koechlin
Charles Wuorinen
Colin Matthews
David Del Tredici
David Diamond
David Maslanka
Donnacha Dennehy
Einojuhani Rautavaara
Elisabeth Lutyens
Ellen Taaffe Zwilich
Elliott Carter
Ensemble Intercontemporain
Ensemble Recherche
Ernst Krenek
Ernst Toch

Your playlist
Mad Rush
Philip Glass

Currently selected: Antonio Vivaldi
Genres: baroque, classical, classical music, early music, italian baroque, italian classical

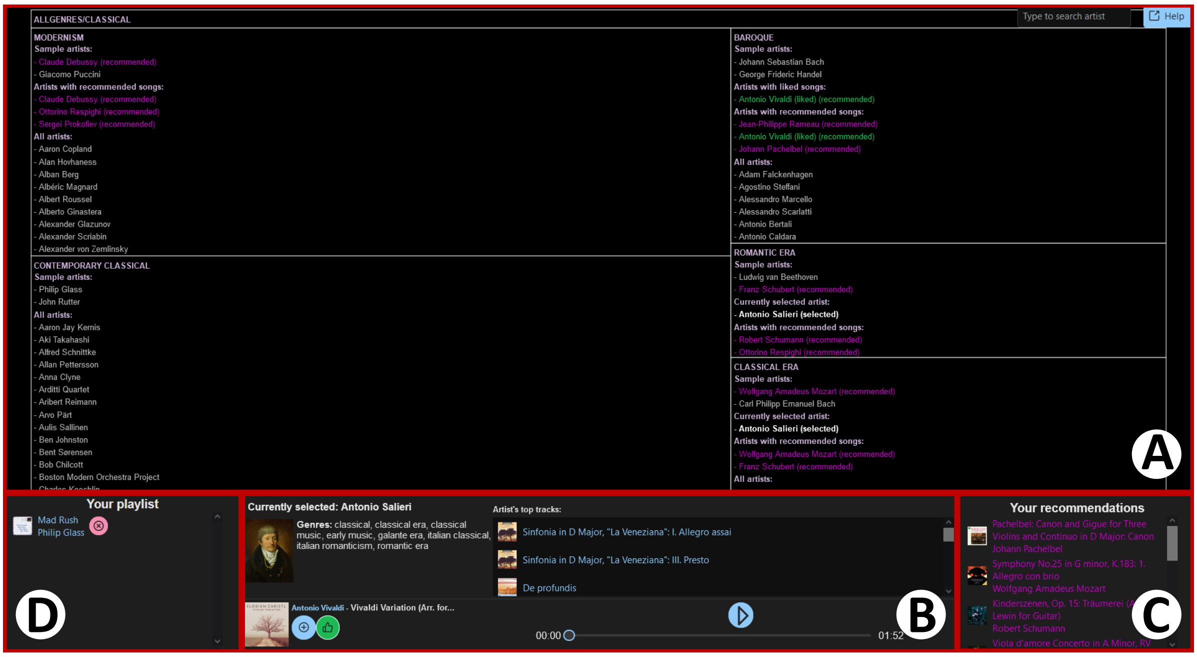
Artist's top tracks:
Vivaldi Variation (Arr. for Piano from Concerto for Strings in G Minor, RV 156) (Liked)
The Four Seasons - Summer in G Minor, RV. 315: III. Presto
The Four Seasons - Winter in F Minor, RV. 297: I. Allegro non molto

Antonio Vivaldi - Vivaldi Variation (Arr. for...
00:00 01:52

Your recommendations
Kinderszenen, Op. 15: No. 7 in F Major, Träumerei
Robert Schumann
Gymnopédie No. 1
Erik Satie
Fauré: Pavane, Op. 50
Gabriel Fauré
Schwanengesang, D. 957: Ständche
"Leise flehen meine Lieder"
Franz Schubert

Figure 3.7: Screenshot of the *MusicExplorationApp* in the *List* condition. In the exploration area (A), the dataset of music artists and songs is visualized as hierarchical list. Currently, the list of the genre *classical* is expanded, showing its subgenres of which *contemporary classical* is expanded. Two representative artists are displayed for each genre. Next to these representative artists, artists with songs previously liked by the active user are displayed in green, and artists with recommended songs are displayed in magenta. The recommended songs are also presented as list in the recommendation area (C). The artist *Antonio Vivaldi* is currently selected and thus appears in the details area (B) along with his 10 most popular songs. Users can like songs here, which results in a recalculation of recommendations. As in the exploration area, liked songs are displayed in green. One song has been added to the current user's playlist, which is displayed in the playlist area (D).

The three conditions of the *MusicExplorationApp* are all based on this dataset, but visualize it differently: as *List*, as *Treemap*, and as *Map*. The other functions and components (i.e. in the areas (B), (C), and (D) in Figure 3.7, Figure 3.8, and Figure 3.9) are implemented in the exact same way. To visualize the item space in the *List* interface, the top-level genres are depicted as entries in an expandable list (Figure 3.7). Clicking on one of these genres displays a list of subgenres, which in turn can be expanded to display all artists associated with the respective subgenre. These two hierarchical levels of genres are represented as cells in the *Treemap* interface (Figure 3.8). Clicking on a top-level genre expands it to cover the entire screen space, allowing the subgenres to be explored. This view also shows a scrollable list of artists. The size of each cell of a top-level genre or subgenre indicates the proportionate number of artists assigned with this genre. Finally, in the *Map* interface, genres are displayed as labels within a two-dimensional map (Figure 3.9). The font size and the color of these labels indicate whether they are top-level genres or subgenres. At the beginning, this interface shows a map displaying only the top-level genres and their representative artists. Then the user can



ALLGENRES/CLASSICAL

MODERNISM
Sample artists:
- Claude Debussy (recommended)
- Giacomo Puccini
Artists with recommended songs:
- Claude Debussy (recommended)
- Ottorino Respighi (recommended)
- Sergei Prokofiev (recommended)
All artists:
- Aaron Copland
- Alan Hovhaness
- Alban Berg
- Albéric Magnard
- Albert Roussel
- Alberto Ginastera
- Alexander Glazunov
- Alexander Scriabin
- Alexander von Zemlinsky

CONTEMPORARY CLASSICAL
Sample artists:
- Philip Glass
- John Rutter
All artists:
- Aaron Jay Kernis
- Aki Takahashi
- Alfred Schnittke
- Allan Pettersson
- Anna Clyne
- Arditti Quartet
- Aribert Reimann
- Arvo Pärt
- Aulis Sallinen
- Ben Johnston
- Bent Sørensen
- Bob Chilcott
- Boston Modern Orchestra Project
- Charles Koechlin

BAROQUE
Sample artists:
- Johann Sebastian Bach
- George Frideric Handel
Artists with liked songs:
- Antonio Vivaldi (liked) (recommended)
Artists with recommended songs:
- Jean-Philippe Rameau (recommended)
- Antonio Vivaldi (liked) (recommended)
- Johann Pachelbel (recommended)
All artists:
- Adam Falckenhagen
- Agostino Steffani
- Alessandro Marcello
- Alessandro Scarlatti
- Antonio Bertali
- Antonio Caldara

ROMANTIC ERA
Sample artists:
- Ludwig van Beethoven
- Franz Schubert (recommended)
Currently selected artist:
- Antonio Salieri (selected)
Artists with recommended songs:
- Robert Schumann (recommended)
- Ottorino Respighi (recommended)

CLASSICAL ERA
Sample artists:
- Wolfgang Amadeus Mozart (recommended)
- Carl Philipp Emanuel Bach
Currently selected artist:
- Antonio Salieri (selected)
Artists with recommended songs:
- Wolfgang Amadeus Mozart (recommended)
- Franz Schubert (recommended)
All artists:

Your playlist
Mad Rush
Philip Glass

Currently selected: Antonio Salieri
Genres: classical, classical era, classical music, early music, galante era, italian classical, italian romanticism, romantic era

Artist's top tracks:
Sinfonia in D Major, "La Veneziana": I. Allegro assai
Sinfonia in D Major, "La Veneziana": III. Presto
De profundis

Antonio Vivaldi - Vivaldi Variation (Arr. for...
00:00 01:52

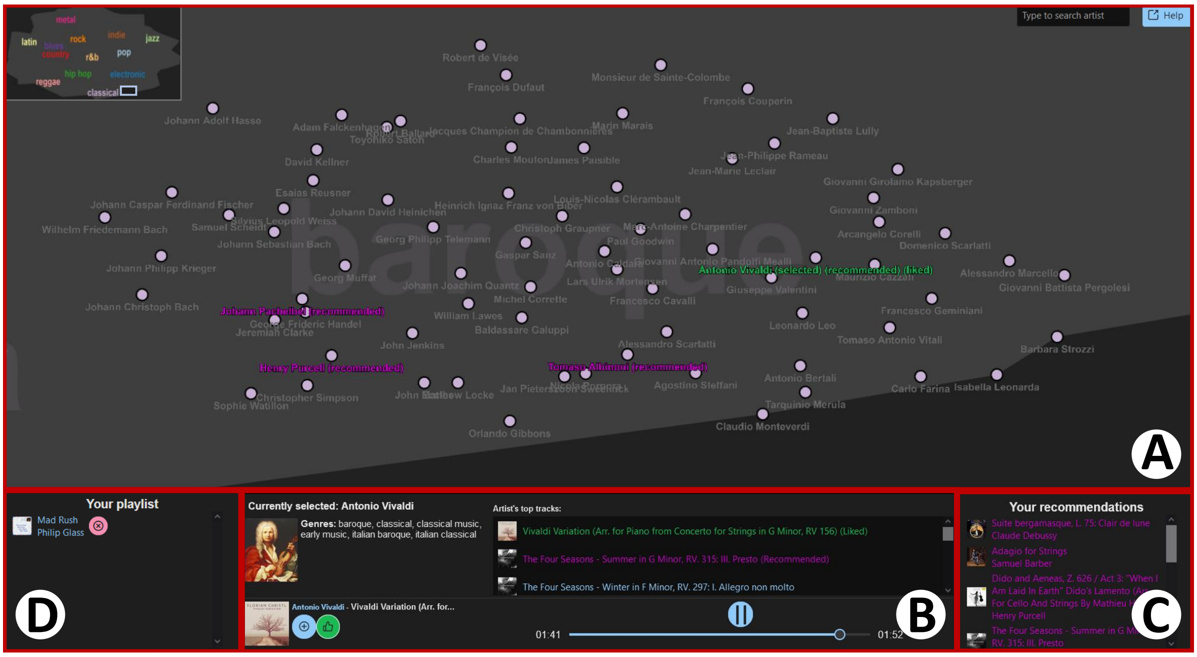
Your recommendations
Pachelbel: Canon and Gigue for Three Violins and Continuo in D Major: Canon
Johann Pachelbel
Symphony No.25 in G minor, K.183: 1. Allegro con brio
Wolfgang Amadeus Mozart
Kinderszenen, Op. 15: Träumerei (A
Lewin for Guitar)
Robert Schumann
Viola d'amore Concerto in A Minor, RV

Figure 3.8: Screenshot of the *MusicExplorationApp* in the *Treemap* condition. In the exploration area (A), the dataset of music artists and songs is visualized as a treemap. Currently, the genre *classical* is shown so that its subgenres can be explored. As with the other conditions, two sample artists (white color) and all artists with liked (green) or recommended (magenta) songs are displayed for each genre. The details area (B), recommendation area (C), and playlist area (D) are implemented in the same way as described for the *List* condition in Figure 3.7.

zoom into the map, revealing the subgenres and eventually all the artists in that area.

Regardless of the item space visualization, users can provide a unary rating for songs (i.e. click a “thumbs up” button). Each time this interaction is performed, new recommendations are requested via the Spotify API and displayed as a list in the recommendation area and in the item space visualization (magenta colored font in the screenshots in Figure 3.7, Figure 3.8, and Figure 3.9).

Using the *MusicExplorationApp*, we conducted a user study to test whether the visualization style influences user-centered qualities of the item space visualization and the underlying RS. More specifically, we asked participants to asses the degree of overview they obtained of the item space in tandem with the perceived recommendation transparency, control, quality, novelty, and variance. Questions for control and overview were created by ourselves, while the other constructs were taken from the inventory of Pu et al. [136]. Additionally, we evaluated the general UX of the different interfaces using the *user experience questionnaire* [147]. During this study, participants had four tasks: (1) rating at least 5 songs; (2) listening to the 10 recommendations they received; (3) creating a playlist of at least 6 songs for themselves; and (4) creating a playlist of at least 6 songs for an evening with friends.



Your playlist

Mad Rush
Philip Glass

Currently selected: Antonio Vivaldi

Genres: baroque, classical, classical music, early music, italian baroque, italian classical

Artist's top tracks:

Vivaldi Variation (Arr. for Piano from Concerto for Strings in G Minor, RV 156) (Liked)

The Four Seasons - Summer in G Minor, RV. 315: III. Presto (Recommended)

The Four Seasons - Winter in F Minor, RV. 297: I. Allegro non molto

Antonio Vivaldi - Vivaldi Variation (Arr. for...

01:41 01:52

Your recommendations

Suite bergamasque, L. 75: Clair de lune
Claude Debussy

Adagio for Strings
Samuel Barber

Dido and Aeneas, Z. 626 / Act 3: "When I Am Laid In Earth" Dido's Lamento (Ar... For Cello And Strings By Mathieu H...
Henry Purcell

The Four Seasons - Summer in G Mi... RV. 315: III. Presto

Figure 3.9: Screenshot of the *MusicExplorationApp* in the *Map* condition. In the exploration area (A), the dataset of music artists and songs is visualized as an item space map. The view is currently zoomed in on the *Baroque* subgenre. A minimap in the upper left corner indicates the position and area of the current view (highlighted with a light blue border). As with the other conditions, artists with liked songs are displayed in green and artists with recommended songs are displayed in magenta. The details area (B), recommendation area (C), and playlist area (D) are implemented in the same way as described for the *List* condition in Figure 3.7.

Results In this study ($N = 91$), we found that the *Map* variant resulted in significantly higher *hedonic* UX compared to the other two conditions. We also observed increased values for the novelty of recommendations participants perceived in the *Map* condition compared to the *Treemap* condition. We note that this perceived novelty was not due to an objective novelty (measured by the recommendations' *self-information* [169, 178]) of recommendations, which was at the same level for all three *MusicExplorationApp* variants. To further unravel the relationship between the condition, the UX, and the recommendation novelty, we conducted a mediation analysis, which revealed that the perceived novelty of the recommendations was fully mediated by the *hedonic* UX. In other words, users believed that their recommendations were more novel because they experienced the *Map* visualization to be innovative and leading edge, when in fact they received the same recommendations as in the other conditions.

Contributions Contrary to our expectations, we did not find any differences between the three variants of the *MusicExplorationApp* in terms of perceived overview of the item space and transparency of recommendations. Thus, one contribution of our work is that apparently very simple means of enabling the exploration of the entire dataset on which an RS relies can help address these qualities. This is particularly true for the perceived recommendation transparency, which was at a very high level in all three conditions. Another key contribution is that relationships between ostensibly unrelated variables should not be underestimated. In this regard, the UX of an application can obviously have side effects on how recommendations are perceived, in this case, on how novel users perceive them to be.

4 Discussion

The previous chapter outlined the research in the papers included in the cumulus of this dissertation. In the present chapter, the findings of this research are discussed in terms of how they answer the RQs introduced in Section 1.1 and what contribution they can make to future research in the respective area.

4.1 Mental Models of Recommender Systems

The first two RQs that underlie my research and that will be discussed here first are: "*Which mental models do users develop of an RS?*" and "*What is the relationship between mental models and users' perception of RSs?*" To answer these questions, I was substantially involved in the two experiments presented in Section 3.1 and Section 3.2. Both aim to investigate the mental models that users construct of the RSs they encounter in their daily lives.

One of the key findings in regards of RQ1 is the general representation of the recommending process (see Figure 3.1). The four steps of this process—data acquisition, inference of user profiles, comparison of user profiles or items, and generation of recommendations—were repeated by all participants of our initial study [127]. Although these results are not generalizable per se due to the small sample size, they provide a rough orientation of how users' mental models are structured. This is supported by the results we found in our second study and a larger sample [103]. When grouping participants by how they sorted recommendation actions, we found that most adhered to a procedural style, mirroring the procedure previously identified in interview responses. From a practical perspective, RSs designers could use the steps of this process to cognitively anchor explanations for recommendations and, for instance, explain how each of these steps is performed in their particular use case. The concepts we found could also be helpful in this process. For example, it could be explained that recommendations are generated by an item-based method—a concept that many participants knew and used to describe their mental model. In any case, our results show that recommendations should be made more transparent to avoid *mystification*, which is associated to several negative attitudes towards RSs.

Negative or positive attitudes towards RSs were investigated more deeply in our second study about mental models [103]. By clustering participants according to their mental models, we analyzed how these were related to the users' perception of RSs. We found the following three major associations: *Concept-based mental models* were found to be connected to low perceived transparency of recommendations. *Procedural mental models*, in contrast, resulted in the highest perceived transparency among the three identified models. Participants with mental models following this style were highly confident that their mental model is correct.

Finally, *socially-focused mental models* were associated with high perceptions of social presence and trustworthiness of RSs.

As a consequence of this tight connection between users' mental models and the way they perceive the system, we argue that the research of RSs should shift from a system-based to a mental-model-based perspective, which I emphasize again in this thesis. Prior research on mental models of intelligent systems, combined with the research I have contributed to as part of this thesis, demonstrate the pivotal role that mental models play in how these systems are perceived and hence how users interact with them. Consequently, research and development of intelligent systems should focus on the elicitation and implications of such models.

As a starting point for such future research, I have formulated three research questions, which are discussed below in tandem with some initial thoughts on how to answer them.

What are the reasons for the diversity of mental models? Although we found some basic similarities in the participants' mental models, overall they were very diverse. The origin of this diversity remains vague, though. Based on our research, we have reason to assume that the specific RS a user thinks of when reporting their mental model is *not* the reason for this diversity. It is more likely that users construct a more general mental model based on a mixture of their previous experience with all RSs they encountered so far. This model is then used every time they interact with an RS—regardless of whether it is on Netflix, Amazon, or YouTube. This is consistent with prior research on mental models, which found that users reuse models they developed with another system [128, 129]. In this sense, the heterogeneity of mental models we identified is due to the different experiences users have had with all the RSs they have encountered, and probably also to the different experiences they have had with one and the same RS.

Another factor that may play a role in the system-independent heterogeneity of mental models is the individual personality of the users. In the experiment, presented and discussed in Section 3.6 and the corresponding paper [97], we applied hierarchical clustering to interaction sequences found while three groups, all exposed to a different application, constructed song playlists. The resulting clustering divided users into groups that were, to some degree, independent of the application to which users were exposed. One can assume that the differences in interaction patterns were due to different mental models that the users held. One reason for why users developed different mental models could lie in their individual user characteristics: cluster groups differed in users' *visual memory capacity* and in their score for the *Big Five* trait *agreeableness*. Although such user characteristics may certainly play a role in why users constructed different mental models, I conclude that further research needs to be conducted in this area. For example, it should be investigated which elements of a mental model are affected by which user characteristics. However, in order to study such connections, a valid instrument for eliciting a user's mental model should first be developed.

How can mental models be elicited? As mentioned earlier, in [97] we analyzed interaction patterns that identified different groups with presumably different mental models. Such pattern analysis could thus perhaps form the basis for capturing mental models of users at runtime. I acknowledge, however, that this is speculative and it would require a considerable amount of research to transform it into a reliable instrument for capturing individual mental models

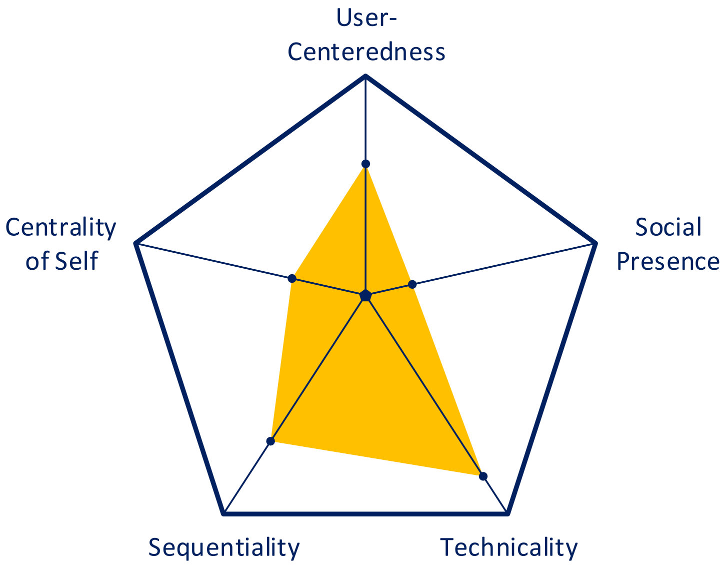


Figure 4.1: A radar chart that could be used to represent mental models: The user holds a rather *user-centered* model, but perceives a very low level of *social presence* in the RS. The user's mental model is highly *technical* and follows a *sequential* structure. Finally, the user does not see their own *self* as very *central* to the system functionality.

through interaction data alone. Card sorting, on the other hand, is a well-established tool for quantitatively identifying mental models, which we combined with hierarchical clustering [103]. Compared to the analysis of interaction patterns, the card-sorting results have the advantage that they allow to investigate which concepts the mental models contain, thus describing the mental models themselves rather than how they affect the interaction. Yet, card sorting is quite time-consuming and thus not applicable in some situations—for instance, when each user's mental model should be elicited before they start using a system. For such situations, a questionnaire-based survey would be more appropriate.

To construct such an instrument, I suggest that the concepts that emerged from our two previous studies could constitute a basis. These five concepts could be assessed using typical questionnaire scales, yielding a personal "*mental model pentagon*" of RSs for each user. A mental model obtained this way might be characterized similar to the radar chart depicted in Figure 4.1. Some of the five concepts in this representation can be measured rather easily. For example, there is already an established questionnaire for measuring *social presence* that can be used more or less immediately [48]. Inventories for other dimensions would require some further development, though. Exploratory factor analysis could help identify items that could be used in such a novel questionnaire-based instrument. The supplementary material of [103] could be used as starting point for collecting candidate items. For example, we have begun to create a scale to measure the dimension of *technicality* as opposed to a *metaphorical* perspective on RSs.

If designed accordingly, an instrument as described above could be used to elicit the mental model of each user (e.g. when registering for an RS) and even repeated later. This would enable a system to adapt to the corresponding mental model of a user.

How should RSs adapt to the mental models of their users? One of our findings is that the comprehension of RSs is generally vague, incomplete, and only partially correct. While technical incorrectness is to a certain degree an inherent characteristic of mental models [128], it was found to cause complications in human-computer interaction [109, 121]—among them the aformentioned *gulfs* of *evaluation* and *execution* [129]. Countermeasures, e.g. in form of textual explanations, have shown potential to narrow such gulfs down but it was proposed that they might not reveal the actual functioning of an RS [160]. To avoid large gulfs and thus false assumptions and erroneous user behavior, we argued that such *ostensible* transparency should be used very carefully to avoid false assumptions and erroneous behavior [103]. Instead systems should provide "real" transparency to their users, i.e. explain how they actually work. This would also help to correct flawed and incomplete mental models of users. The elements that explain the system to the user and thus provide "real" transparency could be personalized if each user's mental model can be elicited at runtime, for instance, by the means discussed above.

One aspect that could be personalized in this context is the *number* of explanatory components. If the model already matches the actual functioning of an RS, the UI elements explaining the recommendations to the user could be reduced to a minimum. On the other hand, if the model is fundamentally different from the algorithmic functioning, more elements could appear to explain how the recommendations are generated. Such an adaptive interface could make more efficient use of the limited available screen space by not explaining something the user is already aware of. In terms of the five dimensions that I propose to capture a user's mental model, system designers could define their RS using the same dimensions. Each user's deviation from this model could then be quantified by matching their elicited model to the system model. Based on the magnitude of this deviation, it would then be possible to automatically determine the extent of the explanatory components.

Aside from the number of explanatory components, a system could also customize *how* these elements explain recommendations to the user. In case, for instance, a user's model scores high on the *metaphorical* dimension, textual explanations could be used that explain recommendations in a metaphorical language. If a user has a more *technical* mental model, the explanations could change to technical flow charts, especially if the captured mental model also scores high on the *procedural* dimension (i.e. *sequentiality* in Figure 4.1). In order to achieve such qualitative adaption of explanations, I propose to design components for each of the five dimensions of the *mental model pentagon* described above. This should also be aligned with ongoing research on personalizing the level of detail in explanation interfaces. It has, for instance, been found that different user characteristics may influence a user's preference for the style and complexity of explanations for recommendations [57, 93].

However, while different explanation styles exist, it has rarely been studied *what users expect* from their explanations. Therefore, I initiated and supervised another experiment in which we asked users to create recommendations for others and justify their choices.

4.2 Comprehension of Recommendations through System- vs. User-generated Explanations

In my third RQ, I ask: "*How do typical system-generated explanations compare to explanations based on natural language created by humans?*". To answer this question and to gain insights into what users deem important when explaining recommendations, we designed an experiment in which users were asked to recommend movies to other users and to explain this recommendation (Section 3.3). We compared these user-generated explanations to system-generated explanations, which followed a style typical for explanations of RSs and were based on similarities between movies. The result was that the user-generated explanations were perceived to be of higher quality, which, interestingly, also affected the perceived quality of recommendations: while the RS generally generated better (i.e. more accurate) recommendations, the higher quality of user-generated explanations compensated for their inferior recommendation accuracy. By applying a counterfactual analysis, we further quantified this effect and found that perceived recommendation quality would increase by half a star on a five-star rating scale if RSs achieved the same explanation quality as users.

One aspect we noticed when manually reviewing the user-generated explanations was that they often adhered to the similarity-based style of typical system-generated explanations. The similarity of recommendations to previously liked items seems to be a natural way of approaching a meaningful explanation. However, the explanations formulated by users went beyond a superficial similarity based on a single item. Participants in our experiment frequently provided a *larger context* of the recommendation. An example for such an explanation is: "*I chose this movie because it is full of action, but also funny, like* Man in Black *or* Pirates of the Caribbean." Apparently, *comprehending* a recommendation includes how it relates to more than one previously highly rated item. This particular explanation also shows how the user defined a broader area in terms of movie genres, in which the rated movies serve as examples.

This observation mirrors to some extent the three levels of explanations provided by Tintarev and Masthoff [160]—but from an item- rather than a user-based perspective. Explanations of our system, which are based on the similarity to an item for which a user has previously expressed a preference explain recommendations on an individual item base. In contrast, many participants in our experiment used explanations on the second level, *contextualizing* the similarity between the recommended and previously rated items, for example, by referring to multiple items at once. The third level, *self-actualization*, goes beyond this. RSs that provide recommendations on this level help users go beyond just receiving recommendations. Instead, they pursue a number of other goals, such as to widen their users' horizon and let them discover topics of the item space they had not previously considered. It remains questionable whether such a level of support for self-actualization can be achieved with textual explanations alone. Even with the given advances in natural language processing, I argue that generating textual explanations can only *mimic* the ability of humans to explain recommendations, but neglects the many additional possibilities computers have to make recommendations more comprehensible and support users' self-actualization. One of these possibilities is the superior ability of computer systems to generate complex yet easy to understand visualizations. In this sense, I have contributed to the development of some systems that provide a *global* context of recommendations—i.e. showing how they relate to all items and content areas in a given repository—and support exploration of the item space, thus aiding users in a range of *self-actualization* goals.

4.3 Global Comprehension of Recommendations by Item Space Visualizations

Providing users with an interface through which they can explore the entire item space would not only allow them to obtain a *global* comprehension of the database and the recommendation context, but also a number of other advantages. Inter alia, this can help avoid *filter bubbles* [102, 131], increase choice confidence [71], explore and develop new preferences [87, 135], and make RSs fairer and more ethical [14, 28].

As a consequence of the above and due to the potential of InfoVis methods to visualize large amounts of information, I formulated two further RQs: "*How can the underlying item space of an RS be visualized to users? How does the visualization style influence the users' perception of an RS?*" To answer these questions, I discuss the results of the three experiments presented in Section 3.4, Section 3.5, and Section 3.6.

4.3.1 How can the underlying item space of an RS be visualized to users? How does the visualization style influence the users' perception of an RS?

Prior research has shown that perceptual variables of RSs such as *transparency*, *control*, and *satisfaction* can be addressed by presenting the entire item space and personalized recommendations within [3, 84, 108]. While these approaches demonstrate the general ability to address these aspects through visualizations, we approached the question of how different visualization styles *compare* in this regard [97]. More specifically, we developed and compared three different applications that visualize the same item space of music artists in three different ways: as *List*, as *Treemap*, and as *Map*. As a result, we found that the *Map* interface was able to display up to about seven times more items simultaneously without sacrificing UX. In fact, the map interface even yielded the highest scores for *hedonic* UX.

I attribute the observation that *hedonic* UX was higher in the *Map* condition than in the *Treemap* condition in part to the fact that the *Treemap* interface was perceived as more cluttered. This conclusion is based on the scores for *feature congestion* (FC), an instrument used to quantify clutter, which were substantially higher in the *Treemap* interface. This measurement is thus a promising tool to predict UX to some extent. We found that a FC of about 2.0 seems reasonable in this regard. A practical conclusion is that maps can display more objects simultaneously without increasing visual clutter.

With respect to the perceived degree of transparency and control, we found that recommendations in all variants were assessed as transparent and to a slightly lesser extent controllable. Apparently, the InfoVis techniques in the *Treemap* and *Map* condition did not add much to what the *List* interface already offered in terms of recommendation *transparency* and *control.* Thus, a general conclusion of this experiment is that even simple means of displaying the item space in the context of RSs can unlock desirable user-centered features of these systems. However, the results of the study with *NewsViz* showed a difference between the visualizations compared. In particular, with *NewsViz*, we were able to demonstrate that a treemap can convey to the user how different information sources contribute to their personalized newsfeed better than traditional sliders. This type of visualization is therefore particularly useful for making users aware of potential biases and thereby preventing filter bubble effects. By extending treemaps to allow users to interactively adjust the cell sizes, we were also able to provide users with more control over the composition of their newsfeed compared to using sliders.

While in *NewsViz* the cell sizes of the treemap were used to display the composition of the active user's preference profile, in the *Treemap* condition of the *MusicExplorationApp* they indicated the popularity of each genre, which is an attribute that is less critical for the control of recommendations. I thus consider a more meaningful use of cell sizes to be a reason for why *NewsViz* was able to unfold its potential better than the *Treemap* interface of the *MusicExplorationApp*: One of the core feature of treemaps—the comparison of proportional arrangement of cell sizes—was deeply integrated into the functioning of the RS in *NewsViz* but not in the *MusicExplorationApp*. It would therefore be interesting to integrate the elicitation of user preferences comparatively deeply into the functionality of the *MusicExplorationApp*.

Such a deep integration of control in the functioning of the *Treemap* condition could in general be realized in a manner analogous to *NewsViz*: by interactively resizing the treemap cells. Users could express their interest in a music genre by resizing the corresponding cell of this genre. In turn, they could then observe their recommendations to change towards that genre. However, this interaction concept would only work on a fairly coarse level, and it would not be trivial to transfer it to an item level. Due to space limitations, artist are currently displayed as list in the *Treemap* condition of the *MusicExplorationApp* and thus cannot be easily manipulated by resizing the treemap cells. One solution would be to cluster the artists within each subgenre to introduce a third, artificial level of hierarchy. To achieve a similar behavior in the *List* condition of the *MusicExplorationApp*, sliders could help to express preference weights for genres, subgenres, and artists. In addition, a third hierarchy level could be introduced, analogous to the *Treemap* condition.

On the *Map* condition, in contrast, the interaction concept described above could be applied comparatively easily. In the *MovieLandscape* application, we have already demonstrated how this could be implemented. This could be transferred to the *Map* condition of the *MusicExplorationApp*, either by changing "landscape heights" precisely as in *MovieLandscape*, or by creating a heatmap to express areas that a user likes or dislikes. Of these two options, I deem the heat map version to be more promising. Even though the landscape metaphor matches well with the concept of a geographic map, visualizing three-dimensional objects or scenes on a two-dimensional screen produces interaction overhead—for example, to change the camera perspective.

Even though the metaphor of a heat map may be more efficient than a three-dimensional landscape, we were able to show with the *MovieLandscape* application that sculpting the landscape works well to express preferences in RSs [98]. With MF, an embedding is learned that is capable of providing recommendations with high accuracy. At the same time, it serves as foundation for the item space visualization, for which the dimensionality is further reduced. This creates a global environment in which recommendations are naturally embedded. As such, I deem *MovieLandscape* as a practical example of one of the implications of our research on mental models: we have formulated that researchers and system developers should "dare to provide

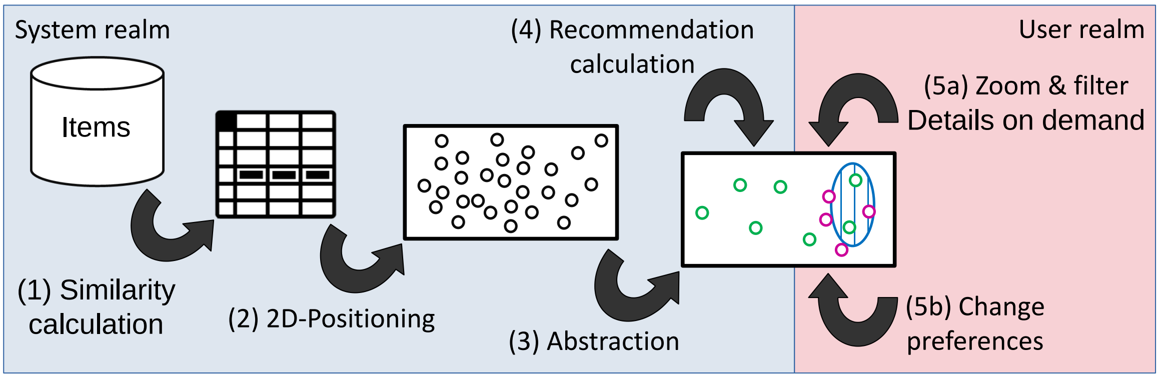


Figure 4.2: The phases for generating the map visualizations based on experiences I made with the *MovieLandscape* [98] and the *MusicExplorationApp* [97]: Similarities between all items of a given repository are computed (1), which are then distributed on a two-dimensional plane according to these similarities (2). Since this distribution would overwhelm users in most cases, it is further abstracted (3), e.g. by clustering and selecting representative examples for each cluster. The resulting visualization can be presented to the user. If the preferences of the active user are known, recommendations can be calculated (4) and displayed within the map visualization. The user can now actively explore the item space (5a) by means of the *information seeking mantra* [151] and express preferences (5b), which in turn trigger a recalculation of recommendations. This process is based on the *InfoVis reference model* presented by Card et al. [16].

real transparency" of their RSs [103]. Due to the consistency of distributing items on the map and computation of recommendations, users are provided with such "real transparency".

I consider this to be closely related to the results of our experiment with explanations generated by users and an RS (Section 3.3): The textual explanations based on CF similarity were not perceived as very helpful and comprehensible. In contrast, the similarities with the *MovieLandscape* were understood, made recommendations transparent, and helped users to express their preferences. Thus, the only apparent difference is that the relationship between preference and recommendations is presented in context of the entire item space. Or, to put it in another way, with the global presentation of the item space and the preferences and recommendations it contains, we have achieved a holistic "contextualization" as also observed in the user-generated explanations.

In the following, I present a general procedure for creating item space maps abstracted from my experiences made with the *MovieLandscape* [98], the *MusicExplorationApp* [97], and other unpublished experiments.

4.3.2 A General Process to Create Item Space Maps

To design item space maps, I propose a process which is displayed in Figure 4.2. This abstract process, which is based on the *InfoVis reference model* presented by Card et al. [16], served as blueprint for the *MovieLandscape* application and the map interface in the *MusicExplorationApp*. First, similarities between all items in a given database are calculated. The type of data used for this is deliberately unspecified here and depends on the particular context, e.g. what data are available. In *MovieLandscape*, we used rating data as is common in RSs based on CF. For the *Map* condition in the *MusicExplorationApp*, on the other hand, we calculated similarities based on music genres and musical features of songs. While this decouples the similarity calculation from the actual recommending algorithm to some extent, we aimed at using aspects of the recommendation algorithm for the similarity calculation to create the "real transparency" mentioned earlier. In this experiment, however, the exact functioning of the recommendation algorithm was inaccessible to us, and we demonstrated how to "graft" the map visualization onto an existing, inaccessible RS.

Based on the resulting similarities, the positioning is performed. Most common *dimensionality reduction* (DR) algorithms can be used for this, as they usually expect similarities[4] as input or can be set up accordingly to accept similarities as input. To decide on a particular algorithm, I suggest measuring the corresponding *continuity* and *trustworthiness* scores [79], which indicate how well the algorithm maintains the structure within the similarities between items. With this similarity distribution, the resulting map employs the *first law of geography* [162], which means that proximity in the map represents the similarity of the corresponding items.

The resulting distribution of items could already be presented to users, but I propose to abstract the item space further first. Again, this can be performed in a variety of ways and depends in part on the data chosen to calculate the similarities. In *MovieLandscape*, for instance, we did not use content data to compute the similarities, so we decided to reduce the number of visible items by selecting representative samples for areas of the map. In the *Map* condition of the *MusicExplorationApp*, on the other hand, the similarity calculation was based on the music genres associated with artists, and we were thus able to add the labels of these genres in the region where the corresponding artists were placed. When no such "natural" groups of items exist, (hierarchical) clustering can be applied to identify regions with similar items that can be abstracted.

After the previous step, the item space map is ready to be presented to users. They should then be provided with interaction tools to explore the item map according to the *information seeking mantra* of Shneiderman [151]: *overview first, zoom and filter, then details-on-demand.* Especially the interaction pattern of zooming should be implemented, as it is a natural and intuitive way of interacting with maps. More precisely, I suggest to implement *semantic zooming* [134], which combines filtering and zooming. Unlike geographic zooming, where elements get progressively larger while zooming in (e.g. when zooming into a particular part of a digital picture), semantic zooming adds further elements on zooming in. To apply this technique, a hierarchical dataset should be prepared. In our experiments, we manually created a hierarchy of genres for the *Map* condition of the *MusicExplorationApp*, and employed clustering in the *MovieLandscape*, where no hierarchical data structure was available beforehand. The third part of the *information seeking mantra* can be achieved by allowing users to click or hover over individual items to retrieve details either in place or in a designated area of the screen.

Finally, users should be able to change their preference profile to exercise control over their

[4] To be more precise, DR algorithms are often performed on *dissimilarities*. However, similarities can easily be converted to dissimilarities and vice versa.

recommendations. In the past, we have implemented two methods that span a continuum of how to achieve control over recommendations: In the *MusicExplorationApp*, users expressed their preferences *discretely* on a per-item base. In *MovieLandscape*, on the other hand, users expressed their preferences *continuously* by lowering or elevating the landscape of entire areas. As discussed earlier, I deem the latter method more appropriate because it integrates the expression of preferences more deeply into the application.

4.4 A Holistic Perspective on Recommender Systems

The research presented in this cumulus has revealed close relationships between conceptional variables in RSs. The perception of *recommendation quality* depends, for instance, not only on how accurately an RS can predict a user's preferences and thus how precisely recommendations are tailored to the user's tastes. Rather, the perceived recommendation quality depends on the *overview* that the user can obtain of the entire item space [102] and on a system's ability to *explain* its recommendations [101]. These two variables have also been shown to be highly influential with respect to other aspects. In addition to recommendation quality, we found that the overview of all items available in an RS influenced perceived *transparency* and, more surprisingly, *system effectiveness*, which is closely related to recommendation quality, but more in terms of general system quality or usability. The explanation quality, on the other hand, affected *social presence* and the *trust* a user puts in the RS, in addition to *recommendation quality.*

In addition to *recommendation quality*, we found that *explanation quality* also influenced the perceived *movie quality*, which led us to assume the difference between recommendation and movie quality was not correctly understood by users [101]. However, I would now like to propose a second interpretation. Like the *halo effect* we alluded to in the experiment with the *MusicExplorationApp* and which is discussed further down below, the *explanation quality* might have in fact affected the subjectively perceived quality of the movie being recommended. While I acknowledge that further experiments are needed to unravel such intricate relationships, I interpret this observation to mean that the context (in this case, an explanation) given to a presented item (in this case, a movie being recommended by an RS) might indeed influence its perceived quality—even though both variables seem isolated at first glance.

Such tightly coupled perceptional variables indicate how subjectively users experience RSs. Consequently, I argue that a user-centered perspective should be substantially more emphasized when studying, developing, and evaluating RSs. One topic that has been underrepresented in the related field is the UX of RSs.

4.4.1 The User Experience of Recommender Systems

For a long time, the focus in the research and development of RSs was mainly on technical and algorithmic aspects, most prominently on their prediction *accuracy.* This was later criticized and means were demanded to assess the quality of an RS *beyond* its accuracy. As a result, the research community of RSs has started to investigate user-centered dimensions such as transparency, user control, and recommendation novelty for evaluating RSs. So far, however, the main focus has been on so-called *do-goals.* In *self-regulation theory* [17], these do-goals are subordinate goals in the overall goal hierarchy of human task completion. Typical goals pursued by RSs, such as *find good item*, *make good decisions*, or *find group consensus* are do-goals: they describe lower-level goals—something a person wants to *do*. Hierarchically above such do-goals are *be-goals*. They pertain to how a person wants to *be*. In RSs, such goals are rarely considered. Examples would be: *be happy*, *be excited*, *feel healthy*, *have fun*. This perspective is often considered in UX design [63, 64], but not in RS development. I suggest that in the future, these two disciplines should be more united and research in RSs should more actively consider perspectives from UX design. As system designers and researchers, we should ask ourselves more experience-based questions when developing RSs. Examples include: "What is it that users enjoy when receiving recommendations?" and "What are the aspects that are exciting and fun in situations as recommendation receiver?"

Adopting such a perspective could also help explain the effect of *hedonic user experience* on *recommendation novelty* that we observed in context of the *MusicExplorationApp*: The interface of the *Map* condition was experienced as more appealing and exciting, motivating a different mindset and thus different *be-goals*. As consequence, users internally switched to another set of lower-level do-goals, in this case: looking for something new. This in turn made them adopt a different perspective on recommendations. While the recommendation algorithm and the objective novelty as calculated by the *self-information* of recommendations were similar across conditions, users presented with the *Map* experienced a higher novelty simply because their *be-goal* was to explore something new and therefore to actively seek out novel items in their recommendations.

As a result of the above, I argue that more attention should be paid to such be-goals in future research of RSs. Since this area is still very unexplored, I suggest conducting qualitative studies to shed some initial light on what such be-goals might be. For example, in the domain of music, it has been found that music is closely related to a person's identity and sense of being unique [144]. Consequently, this feeling helps to differentiate oneself from others, but also to connect peers in social groups. Thus, a be-goal of a music RS would be to help users better understand their own interests, but also the musical interests of friends. In this process, the role of an RS would be considerably different from the typical presentation of top-n recommendations. Rather, an RS would need to be more a tool that emphasizes where a user's interests lie and how they relate to thus far unexplored areas and to the preferences of the user's social group.

4.5 Limitations

There are some limitations with respect to the research conducted in the context of this thesis. One pertains to the domains tested. Mostly I concentrated on domains in the entertainment sector (i.e. music and movies) but also news in the study with *NewsViz*. Some reasons for this decision are that well-established datasets in some of these domains exist (e.g. *Movielens*[5]), that many applications in these domains make visible use of RSs which many users are familiar with and it is thus comparably easy to find participants for studies, and that items in these domains can be consumed during a system's evaluation, which is necessary for many domains to assess the quality of the items and recommendations [107]. Apart from that, I deliberately chose *experience* rather than *search* products as subject to recommendations as a domain in which

[5] https://grouplens.org/datasets/movielens/

RSs are particularly relevant [149, 176]. However, due to the restriction on these domains, one limitation of my findings is that it cannot be said with certainty how they are generalizable for other domains. It would be particularly interesting to test domains with higher risk involved in future research. Especially with respect to the trust a users puts in an RS and the decision confidence probably emitted by comprehending how the recommendations relate to the entire item space, e.g. in terms of coverage, this would represent a valuable extension to my work presented here.

With respect to the transparency that is conveyed by presenting users with an explorable depiction of the entire item space, I can only report high values for all three interfaces in the *MusicExplorationApp* without being able to compare them to a baseline without such means to explore the item space. In this experiment, I chose to use a simple depiction of the underlying dataset as list to create a baseline, the experimental conditions with map and treemap are compared to. This decision was made due to fairness concerns—a aspect that some other related works fall short at. Yet, such other works have already shown that such depictions increase the transparency and I am thus confident that the conditions in *MusicExplorationApp* would too. Nonetheless, a future experiment could run a comparison of the existing conditions with a baseline in which the item space depicting components are masked and in which users need to use the search bar to explore the item space.

A third limitation is that the studies in [98], [102], and [97], which all rely on rather complex visualizations were performed on desktop computers and thus on relatively large screens. While currently, online users still mainly use desktop computers or laptops, the trend goes into direction of mobile devices [13]. A future direction of research should thus be to investigate how the benefits of InfoVis can be used on mobile devices. In terms of item space maps, one way to achieve this would be to present the space in terms of a “ego perspective” [69]. This, however, would sacrifice the larger overview and is thus probably not very promising. A more direct translation would be to introduce another level of hierarchy with even less top-level landmarks (e.g. top-level music genres). It needs to be evaluated if this diminishes the perceived overview, though. The same accounts for the computational power of the device the user is using. Especially the visualizations oftentimes need some memory to work properly. This might be overcome in re-programming the applications, for instance, by rendering pictures of the current view on the server-side. This could, however, result in some other ramifications, e.g. in regards of increased network traffic. Perhaps some intelligent method for switching between client-side and server-side computation could be a solution.

5 Conclusions and Future Work

Researchers and developers of RSs and intelligent systems have begun to recognize the pivotal role of user-centered aspects that go beyond the technical precision of the underlying algorithms. One approach is to open the *black box* and make RSs more transparent, for example by providing textual explanations—a method employed not only in experimental research but also in industry. While this undoubtedly advances intelligent systems in a more user-friendly direction, I argue in this dissertation that this should only be considered a first step, and that we have merely begun to understand what a *deep* integration of the interrelated disciplines of human-computer interaction, behavioral psychology, and artificial intelligence means for the development of next-generation intelligent systems.

One of the contributions I make in this regard through my research presented here is to uncover the mental models users of RS develop and to make the implications for the development of such systems explicit. In two user studies, we revealed a wide range of different conceptions of how RS work. A closer look at the mental models found showed that the majority of them were rather technical in nature. For example, the concepts of item-based and user-based CF were clearly evident in some mental models. On the other hand, a considerable number of participants expressed high levels of uncertainty and mystified RSs, which in some cases led them to reject their use entirely—probably due to a lack of trust. To address this, I suggest assisting users in creating a more accurate mental model and correcting it if it is erroneous. Our results show that users do not only *want* to develop such a more accurate mental model, but also that they are *confident* to be able to understand RSs if they were explained to them.

In another study, we investigated whether typically used textual explanations can help users understand RSs and thus trust them more. The results suggest that especially similarity-based explanations of contemporary RSs are too superficial to be trusted by human recipients. However, if these systems were capable of explaining recommendations in a similar fashion to humans, they could not only instill trust in users for their suggestions, but their recommendation accuracy would also be perceived as higher. The direction of automatically generating textual explanations that resemble those of humans is difficult, perhaps even impossible, to achieve. Machines lack the experience, creativity, and empathy that humans possess when explaining the recommendations they make. Although progress is constantly being made in this area, I argue that we should stop trying to mimic a human way of making transparent recommendations, e.g. through speech-based explanations, and instead focus on the areas where intelligent systems have an advantage over humans. One of these directions, which is more or less completely beyond the human ability to explain recommendations, is the use of the expressive power of InfoVis.

In the research on integrating InfoVis into RSs, we focused the ability of visualizing large data sets of RSs by using maps and treemaps. Displaying the global context in which recommendations are embedded can unlock a comprehension of the bigger picture that is unlikely to be achievable through the textual explanations discussed above. Our experiments have shown, for instance, that a global context can help make users aware of potential biases or blind spots, provide better control over recommendations, and make RSs more fun and exciting to use. While I acknowledge that such global visualizations cannot realistically be displayed in every single situation recommendations are provided to users, I am confident that taking advantage of such representations can have a major impact on how RSs are comprehended and thus used in the future.

Another factor emphasized in the research presented here is that results of user studies with RSs should not be interpreted as isolated. In different experiments, I have observed that user-centered aspects are strongly interconnected and can have unexpected side effects. Examples for such interrelationships include the connection between hedonic UX and the recommendation novelty, the perceived quality of explanations and of recommendations, and the influence of the perceived degree of overview on recommendation quality. Mediation analysis and structural equation modeling are two statistical tools that enable the study of such influences. As a result, future research in RSs should consider these systems more holistically and, for instance, comprehend RSs as more than just the component providing recommendations, but also include the entire ecosystem around it. This includes the means of exploring the item space and context of recommendations, the efficiently expression of preferences, and the general control users have over the recommendation engine.

To integrate all these aspects even deeper in an RS, a concrete next step for future work is a reimplementation of the *MusicExplorationApp*. The concept of controlling preferences and thus recommendations should be incorporated even more integrally into the core functionality of the three visualization variants list, treemap, and map, as discussed in Section 4.3.1. I am confident that this would further increase user utility and likely reveal more differences between the conditions. In this context, another user study should be conducted that also tests a fourth interface without a browsable item space. Consideration should also be given to introducing a version of this application for smaller displays or even mobile devices. However, this would most likely involve a rather drastic change in the conceptual design of the application, as currently the screen space is used quite exhaustively.

In such a next experiment, the relationship between mental models and the perception of the item space visualization could also be investigated more deeply. In this context, a qualitative pre-study could investigate how maps can be leveraged better to anchor a user's mental model with the depiction. For example, the presentation of abstract landmarks (e.g. in the form of tags clouds for regions) for the orientation could be the subject of such an investigation.

Bibliography

[1] Qurat Ul Ain, Mohamed Amine Chatti, Mouadh Guesmi, and Shoeb Joarder. A multi-dimensional conceptualization framework for personalized explanations in recommender systems. In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 11–23, 2022.

[2] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, pages 275–279. ACM, 2016. ISBN 978-1-4503-4368-8. doi: 10.1145/2930238.2930280.

[3] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay: Interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, 121:142–159, 2019. doi: 10.1016/j.ijhcs.2018.04.004.

[4] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. doi: 10.1126/science.aaa1160.

[5] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 333–336. ACM, 2011. ISBN 978-1-4503-0683-6. doi: 10.1145/2043932.2043996.

[6] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. How to recommend?: User trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 287–300. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3025171.3025209.

[7] M. Bilgic and Raymond J. Mooney. Explaining recommendations: satisfaction vs. promotion. In *Proceedings of Beyond Personalization Workshop, IUI*, 2005.

[8] Robert P. Biuk-Aghai, Patrick Cheong-Iao Pang, and Bin Pang. Map-like visualisations vs. treemaps: An experimental comparison. In *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction*, VINCI '17, pages 113–120. ACM, 2017. ISBN 978-1-4503-5292-5. doi: 10.1145/3105971.3105976.

[9] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management*, 49(1):13–33, 2013. doi: 10.1016/j.ipm.2012.06.004.

[10] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark Graus. Understanding choice overload in recommender systems. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 63–70. ACM, 2010. doi: 10.1145/1864708.1864724.

[11] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer New York, 2 edition, 2005.

[12] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 35–42. ACM, 2012. ISBN 978-1-4503-1270-7. doi: 10.1145/2365952.2365964.

[13] Christina Bröhl, Peter Rasche, Janina Jablonski, Sabine Theis, Matthias Wille, and Alexander Mertens. Desktop pc, tablet pc, or smartphone? an analysis of use preferences in daily activities for different technology generations of a worldwide sample. In Jia Zhou and Gavriel Salvendy, editors, *Human Aspects of IT for the Aged Population. Acceptance, Communication and Participation*, pages 3–20. Springer International Publishing, 2018. ISBN 978-3-319-92034-4.

[14] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR, 2018. URL `https://proceedings.mlr.press/v81/burke18a.html`.

[15] André Calero Valdez, Martina Ziefle, and Katrien Verbert. Hci for recommender systems: The past, the present and the future. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 123–126. ACM, 2016. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959158.

[16] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in information visualization: Using vision to think*. The Morgan Kaufmann series in interactive technologies. Morgan Kaufmann Publishers, San Francisco, Calif, 1999. ISBN 1-55860-533-9.

[17] Charles S. Carver and Michael F. Scheier. *On the self-regulation of behavior*. On the self-regulation of behavior. Cambridge University Press, New York, NY, US, 1998. ISBN 0-521-57204-5.

[18] Pablo Castells, Neil Hurley, and Saúl Vargas. Novelty and diversity in recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 603–646. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4.

[19] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. Searchlens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 498–509. ACM, 2019. ISBN 978-1-4503-6272-6. doi: 10.1145/3301275.3302321.

[20] Shuo Chang, F. Maxwell Harper, and Loren Terveen. Using groups of items for preference elicitation in recommender systems. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1258–1269. ACM, 2015. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675210.

[21] Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 175–182. ACM, 2016. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959153.

[22] Chaomei Chen. Information visualization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):387–403, 2010. doi: 10.1002/wics.89.

[23] Li Chen and Ho Keung Tsoi. Users' decision behavior in recommender interfaces: Impact of layout design. In *RecSys' 11 Workshop on Human Decision Making in Recommender Systems*, 2011.

[24] S. Chen, S. Li, and X. Yuan. R-map: A map metaphor for visualizing information reposting process in social media. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1204–1214, 2020. doi: 10.1109/TVCG.2019.2934263.

[25] Jaewon Choi, Hong Joo Lee, and Yong Cheol Kim. The influence of social presence on evaluating personalized recommender systems. In *Pacific Asia Conference on Information Systems*, PACIS, 2009.

[26] Nancy J. Cooke. Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, 41(6):801–849, 1994. doi: 10.1006/ijhc.1994.1083.

[27] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Thousand Oaks, California, 2008.

[28] Pranav Dandekar, Ashish Goel, and David T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15): 5791–5796, 2013. doi: 10.1073/pnas.1217220110.

[29] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3):554–559, 2016. doi: 10.1073/pnas.1517441113.

[30] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 120:1–120:12. ACM, 2018. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173694.

[31] Dominic DiFranzo and Kristine Gloria-Garcia. Filter bubbles and fake news. *XRDS*, 23 (3):32–35, 2017. doi: 10.1145/3055153.

[32] Jérôme Dinet and Muneo Kitajima. “draw me the web”: Impact of mental model of the web on information search performance of young users. In *Proceedings of the 23rd Conference on l’Interaction Homme-Machine*, IHM ’11. Association for Computing Machinery, 2011. ISBN 9781450308229. doi: 10.1145/2044354.2044358.

[33] Weicong Ding, Dinesh Govindaraj, and S. V. N. Vishwanathan. Whole page optimization with global constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19, pages 3153–3161. Association for Computing Machinery, 2019. ISBN 9781450362016. doi: 10.1145/3292500.3330675.

[34] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. Towards understanding latent factors and user profiles by enhancing matrix factorization with tags. In Ido Guy and Amit Sharma, editors, *Poster Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016): Boston, USA, September 17, 2016*, volume 1688, 2016. URL `http://ceur-ws.org/Vol-1688/paper-20.pdf`.

[35] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. Explaining recommendations by means of user reviews. In *Proceedings of the 1st Workshop on Explainable Smart Systems (ExSS ’18)*, 2018. URL `http://ceur-ws.org/Vol-2068/exss8.pdf`.

[36] Fan Du, Sana Malik, Georgios Theocharous, and Eunyee Koh. Personalizable and interactive sequence recommender system. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA ’18, pages 1–6. Association for Computing Machinery, 2018. ISBN 9781450356213. doi: 10.1145/3170427.3188506.

[37] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing transparency design into practice. In *23rd International Conference on Intelligent User Interfaces*, IUI ’18, pages 211–223. ACM, 2018. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172961.

[38] Michael D. Ekstrand and Martijn C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys ’16, pages 221–224. ACM, 2016. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959179.

[39] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 1–14. ACM, 2019. doi: 10.1145/3290605.3300724.

[40] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. A methodology for the offline evaluation of recommender systems in a user interface with multiple carousels. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 10–15. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383677. doi: 10.1145/3450614.3461680.

[41] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. Measuring the user satisfaction in a recommendation interface with multiple carousels. In *ACM International Conference on Interactive Media Experiences*, IMX ’21, pages 212–217. Association for Computing Machinery, 2021. ISBN 9781450383899. doi: 10.1145/3452918.3465493.

[42] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006.

[43] Pavel Gajdusek and Ladislav Peska. Spotifygraph: Visualisation of user's preferences in music. In Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras, editors, *MultiMedia Modeling*, pages 379–384. Springer International Publishing, 2021. ISBN 978-3-030-67835-7.

[44] Emden Gansner, Yifan Hu, and Stephen Kobourov. Gmap: Visualizing graphs and clusters as maps. In *Visualization Symposium, Pacific Asia-Pacific*, pages 201–208, 2010.

[45] Peter Gärdenfors. *Conceptual spaces: The geometry of thought.* MIT Press, 2004.

[46] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014. doi: 10.1016/j.ijhcs.2013.12.007.

[47] David Gefen and Detmar W. Straub. Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, 21(4):389–400, 1997. URL `http://www.jstor.org/stable/249720`.

[48] David Gefen and Detmar W. Straub. Consumer trust in b2c e-commerce and the importance of social presence: experiments in e-products and e-services. *Omega*, 32(6): 407–424, 2004. doi: 10.1016/j.omega.2004.01.006.

[49] Dedre Gentner and Albert L. Stevens, editors. *Mental Models.* Psychology Press, New York, NY, USA, 1st edition edition, 1983. doi: 10.4324/9781315802725.

[50] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12. Association for Computing Machinery, 2020. ISBN 9781450367080. doi: 10.1145/3313831.3376316.

[51] Muheeb Faizan Ghori, Arman Dehpanah, Jonathan Gemmell, Hamed Qahri-Saremi, and Bamshad Mobasher. Does the user have a theory of the recommender? a pilot study. In *Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, IntRS '19, pages 77–85. CEUR-WS.org, 2019.

[52] Justin Scott Giboney, Susan A. Brown, Paul Benjamin Lowry, and Jay F. Nunamaker. User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, 72:1–10, 2015. doi: 10.1016/j.dss.2015.02.005.

[53] Dorota Glowacka, Tuukka Ruotsalo, Ksenia Konuyshkova, kumaripaba Athukorala, Samuel Kaski, and Giulio Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 117–128. ACM, 2013. ISBN 978-1-4503-1965-2. doi: 10.1145/2449396.2449413.

[54] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[55] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. Smallworlds: Visualizing social recommendations. *Computer Graphics Forum*, 29(3):833–842, 2010. doi: 10.1111/j.1467-8659.2009.01679.x.

[56] Mouadh Guesmi, Mohamed Amine Chatti, Alptug Tayyar, Qurat Ul Ain, and Shoeb Joarder. Interactive visualizations of transparent user models for self-actualization: A human-centered design approach. *Multimodal Technologies and Interaction*, 6(6), 2022. doi: 10.3390/mti6060042.

[57] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. Explaining user models with different levels of detail for transparent recommendation: A user study. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 14 pages. ACM, 2022. doi: 10.1145/3511047.3537685.

[58] Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 547–601. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4.

[59] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble? effects of personalization on the diversity of google news. *Digital Journalism*, 6(3):330–343, 2018. doi: 10.1080/21670811.2017.1338145.

[60] Jaron Harambam, Natali Helberger, and Joris van Hoboken. Democratizing algorithmic news recommenders: How to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180088, 2018. doi: 10.1098/rsta.2018.0088.

[61] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pages 69–77. ACM, 2019. ISBN 978-1-4503-6243-6. doi: 10.1145/3298689.3347014.

[62] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 3–10. ACM, 2015. ISBN 978-1-4503-3692-5. doi: 10.1145/2792838.2800179.

[63] Marc Hassenzahl and Noam Tractinsky. User experience - a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006. doi: 10.1080/01449290500330331.

[64] Marc Hassenzahl, Michael Burmester, and Franz Koller. User experience is all there is: Twenty years of designing positive experiences and meaningful technology. *i-com*, 20(3): 197–213, 2021. doi: 10.1515/icom-2021-0034.

[65] Chen He, Denis Parra, and Katrien Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016. doi: 10.1016/j.eswa.2016.02.013.

[66] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Commun. ACM*, 53(6):59–67, 2010.

[67] Natali Helberger. On the democratic role of news recommenders. *Digital Journalism*, 7 (8):993–1012, 2019. doi: 10.1080/21670811.2019.1623700.

[68] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250. ACM, 2000. ISBN 1-58113-222-0. doi: 10.1145/358916.358995.

[69] Sebastian Huber, Markus Schedl, and Peter Knees. Nepdroid: An intelligent mobile music player. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12. Association for Computing Machinery, 2012. ISBN 9781450313292. doi: 10.1145/2324796.2324862.

[70] S. S. Iyengar and Mark R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 76(6):995–1006, 2000. doi: 10.1037/0022-3514.79.6.995.

[71] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 7–10. ACM, 2016. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959186.

[72] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

[73] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 47–51. ACM, 2010. ISBN 978-1-4503-0407-8.

[74] Michael Jugovac and Dietmar Jannach. Interacting with recommenders–overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3): 10:1–10:46, 2017. doi: 10.1145/3001837.

[75] Martijn Kagie, Michiel van Wezel, and Patrick J.F. Groenen. Map based visualization of product catalogs. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 547–576. Springer US, Boston, MA, 2011. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3_17.

[76] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.*, 7(1):2:1–2:42, 2017. doi: 10.1145/2926720.

[77] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 247–251. ACM, 2015. ISBN 978-1-4503-3306-1. doi: 10.1145/2678025.2701371.

[78] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, 2018. doi: 10.1016/j.ipm.2018.04.008.

[79] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4(1):48, 2003. doi: 10.1186/1471-2105-4-48.

[80] Rahul Katarya, Ivy Jain, and Hitesh Hasija. An interactive interface for instilling trust and providing diverse recommendations. In *2014 International Conference on Computer and Communication Technology (ICCCT)*, pages 17–22, 2014. doi: 10.1109/ICCCT.2014.7001463.

[81] Mandy Keck and Dietrich Kammer. Exploring visualization challenges for interactive recommender systems. In *VisBIA 2018*, 2018. URL `http://ceur-ws.org/Vol-2108/paper3.pdf`.

[82] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. doi: 10.1109/2945.981847.

[83] Daniel Kluver, Tien T. Nguyen, Michael D. Ekstrand, Shilad Sen, and John Riedl. How many bits per rating? In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 99–106. ACM, 2012. ISBN 978-1-4503-1270-7.

[84] Peter Knees, Markus Schedl, Tim Pohle, and Gerhard Widmer. An innovative three-dimensional user interface for exploring music collections enriched. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pages 17–24. ACM, 2006. ISBN 1-59593-447-2. doi: 10.1145/1180639.1180652.

[85] Bart P. Knijnenburg, Niels J.M. Reijmer, and Martijn C. Willemsen. Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 141–148. ACM, 2011. ISBN 978-1-4503-0683-6. doi: 10.1145/2043932.2043960.

[86] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):441–504, 2012. doi: 10.1007/s11257-011-9118-4.

[87] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 11–14. ACM, 2016. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959189.

[88] Christie Kodama, Beth St. Jean, Mega Subramaniam, and Natalie Greene Taylor. There's a creepy guy on the other end at google!: engaging middle school students in a drawing activity to elicit their mental models of google. *Information Retrieval Journal*, 20(5):403–432, 2017. doi: 10.1007/s10791-017-9306-x.

[89] Joseph A. Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012. doi: 10.1007/s11257-011-9112-x.

[90] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

[91] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 91–142. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_3.

[92] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 84–88. ACM, 2017. ISBN 978-1-4503-4652-8. doi: 10.1145/3109859.3109915.

[93] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 379–390. ACM, 2019. ISBN 978-1-4503-6272-6. doi: 10.1145/3301275.3302306.

[94] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1–10. ACM, 2012. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2207678.

[95] Jayachithra Kumar and Nava Tintarev. Using visualizations to encourage blind-spot exploration. In *5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, IntRS '18, 2018.

[96] Johannes Kunkel and Jürgen Ziegler. Visualizing item spaces to increase transparency and control in recommender systems. In *AI and HCI Workshop at CHI'19*. 2019.

[97] Johannes Kunkel and Jürgen Ziegler. A comparative study of item space visualizations for recommender systems. *International Journal of Human-Computer Studies*, 172, 2022. doi: https://doi.org/10.1016/j.ijhcs.2022.102987.

[98] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 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, pages 3–15. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3025171.3025189.

[99] Johannes Kunkel, Tim Donkers, Catalin-Mihai Barbu, and Jürgen Ziegler. Trust-related effects of expertise and similarity cues in human-generated recommendations. In *Companion Proceedings of the 23rd International on Intelligent User Interfaces: 2nd Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE)*, 2018. URL `http://ceur-ws.org/Vol-2068/humanize5.pdf`.

[100] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 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)*, 2018. URL `https://arxiv.org/pdf/1808.10260.pdf`.

[101] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12. ACM, 2019. doi: 10.1145/3290605.3300717.

[102] Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. Newsviz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 126–135. ACM, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394869.

[103] Johannes Kunkel, Thao Ngo, Jürgen Ziegler, and Nicole Krämer. Identifying group-specific mental models of recommender systems: A novel quantitative approach. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Proceedings of the IFIP Conference on Human-Computer Interaction*, INTERACT 2021, pages 383–404. Springer International Publishing, 2021. ISBN 978-3-030-85610-6. doi: 10.1007/978-3-030-85610-6_23.

[104] Janice Langan-Fox, Sharon Code, and Kim Langfield-Smith. Team mental models: Techniques, methods, and analytic approaches. *Human Factors*, 42(2):242–271, 2000. doi: 10.1518/001872000779656534.

[105] Benedikt Loepp, Tim Hussein, and Jürgen Ziegler. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd International Conference on Human Factors in Computing Systems*, CHI '14, pages 3085–3094. ACM, 2014. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557069.

[106] Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proceedings of the 33rd ACM International Conference on Human Factors in Computing Systems*, CHI '15. ACM, 2015.

[107] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. Impact of item consumption on assessment of recommendations in user studies. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 49–53. ACM, 2018. ISBN 978-1-4503-5901-6. doi: 10.1145/3240323.3240375.

[108] Boxuan Ma, Min Lu, Yuta Taniguchi, and Shin'ichi Konomi. Courseq: the impact of visual and interactive course recommendation in university environments. *Research and Practice in Technology Enhanced Learning*, 16(1):18, 2021. doi: 10.1186/s41039-021-00167-7.

[109] Stephann Makri, Ann Blandford, Jeremy Gow, Jon Rimmer, Claire Warwick, and George Buchanan. A library or just another information resource? a case study of users' mental models of traditional and digital libraries. *Journal of The American Society for Information Science and Technology*, 58(3):433–445, 2007. doi: 10.1002/asi.20510.

[110] Stephann Makri, Ann Blandford, Mel Woods, Sarah Sharples, and Deborah Maxwell. "making my own luck": Serendipity strategies and how to support them in digital information environments. *Journal of the Association for Information Science and Technology*, 65(11):2179–2194, 2014. doi: 10.1002/asi.23200.

[111] D. Mashima, S. Kobourov, and Y. Hu. Visualizing dynamic data with maps. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1424–1437, 2012.

[112] Lorraine McGinty and James Reilly. On the evolution of critiquing recommenders. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 419–453. Springer US, Boston, MA, 2011. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3_13.

[113] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359, 2002. doi: 10.1287/isre.13.3.334.81.

[114] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In Peter Brusilovsky, Albert Corbett, and Fiorella de Rosis, editors, *User Modeling 2003: 9th International Conference, UM 2003 Johnstown, PA, USA, June 22–26, 2003 Proceedings*, pages 178–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-44963-8. doi: 10.1007/3-540-44963-9_24.

[115] Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1097–1101. ACM, 2006. ISBN 1-59593-298-4. doi: 10.1145/1125451.1125659.

[116] Sean M. McNee, John Riedl, and Joseph A. Konstan. Making recommendations better: An analytic model for human-recommender interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1103–1108. ACM, 2006. ISBN 1-59593-298-4. doi: 10.1145/1125451.1125660.

[117] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc, 2013. URL `https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf`.

[118] Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. Controlling spotify recommendations: Effects of personal characteristics on music recommender user interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, pages 101–109. ACM, 2018. ISBN 978-1-4503-5589-6. doi: 10.1145/3209219.3209223.

[119] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018. doi: 10.1080/1369118X.2018.1444076.

[120] Daniel R. Montello, Sara Irina Fabrikant, Marco Ruocco, and Richard S. Middleton. Testing the first law of cognitive geography on point-display spatializations. In Walter Kuhn, Michael F. Worboys, and Sabine Timpf, editors, *Spatial Information Theory.*

Foundations of Geographic Information Science, pages 316–331. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-39923-0. doi: 10.1007/978-3-540-39923-0_21.

[121] Jack Muramatsu and Wanda Pratt. Transparent queries: Investigation users' mental models of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 217–224. Association for Computing Machinery, 2001. ISBN 1581133316. doi: 10.1145/383952.383991.

[122] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31 (3):274–295, 2014. doi: 10.1007/s00357-014-9161-z.

[123] Cataldo Musto, Marco de Gemmis, Pasquale Lops, Fedelucio Narducci, and Giovanni Semeraro. Semantics and content-based recommendations. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 251–298. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_7.

[124] Sayooran Nagulendra and Julita Vassileva. Understanding and controlling the filter bubble through interactive visualization: A user study. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 107–115. ACM, 2014. ISBN 978-1-4503-2954-5. doi: 10.1145/2631775.2631811.

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

[126] Thao Ngo and Nicole Krämer. It's just a recipe?—comparing expert and lay user understanding of algorithmic systems. *Technology, Mind, and Behavior*, 2(4), 2021. doi: 10.1037/tmb0000045.

[127] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 183–191. ACM, New York, NY, USA, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394841.

[128] Donald A. Norman. Some observations on mental models. In Dedre Gentner and Albert L. Stevens, editors, *Mental Models*, pages 7–14. Psychology Press, New York, NY, USA, 1983.

[129] Donald A. Norman. *The design of everyday things.* Basic Books, Inc., New York, NY, USA, 1988. ISBN 978-0-465-06710-7.

[130] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. Peerchooser: Visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1085–1088. ACM, 2008. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357222.

[131] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* The Penguin Press, New York, NY, USA, 2011. ISBN 1-101-50572-9.

[132] Denis Parra, Peter Brusilovsky, and Christoph Trattner. See what you want to see: visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, IUI '14, pages 235–240, 2014. doi: 10.1145/2557500.2557542.

[133] Alexandre Passant. dbrec — music recommendations using dbpedia. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web – ISWC 2010*, pages 209–224. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-17749-1.

[134] Ken Perlin and David Fox. Pad: An alternative approach to the computer interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 57–64. Association for Computing Machinery, 1993. ISBN 0897916018. doi: 10.1145/166117.166125.

[135] Savvas Petridis, Nediyana Daskalova, Sarah Mennicken, Samuel F. Way, Paul Lamere, and Jennifer Thom. Tastepaths: Enabling deeper exploration and understanding of personal preferences in recommender systems. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pages 120–133. Association for Computing Machinery, 2022. ISBN 9781450391443. doi: 10.1145/3490099.3511156.

[136] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164. ACM, 2011. ISBN 978-1-4503-0683-6. doi: 10.1145/2043932.2043962.

[137] Behnam Rahdari, Branislav Kveton, and Peter Brusilovsky. The magic of carousels: Single vs. multi-list recommender systems. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, pages 166–174. Association for Computing Machinery, 2022. ISBN 9781450392334. doi: 10.1145/3511095.3531278.

[138] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, 1997.

[139] Paul Resnick, R. Kelly Garrett, Travis Kriplean, Sean A. Munson, and Natalie Jomini Stroud. Bursting your (filter) bubble: Strategies for promoting diverse exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, CSCW '13, pages 95–100. ACM, 2013. ISBN 978-1-4503-1332-2. doi: 10.1145/2441955.2441981.

[140] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Introduction and challenges. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 1–34. Springer US, Boston, MA, 2015. ISBN 978-0-387-85819-7.

[141] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook.* Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4.

[142] M. Rossetti, F. Stella, and M. Zanker. Towards explaining latent factors with topic models in collaborative recommender systems. In *2013 24th International Workshop on Database and Expert Systems Applications*, pages 162–167. IEEE Computer Society, 2013. doi: 10.1109/DEXA.2013.26.

[143] David E. Rumelhart and Donald A. Norman. Representation in memory, 1983.

[144] Even Ruud. Music and identity. *Nordisk Tidsskrift for Musikkterapi*, 6(1):3–13, 1997. doi: 10.1080/08098139709477889.

[145] Yuri Saito and Takayuki Itoh. Musicube: A visual music recommendation system featuring interactive evolutionary computing. In *Proceedings of the 2011 Visual Information Communication - International Symposium*, VINCI '11. ACM, 2011. ISBN 978-1-4503-0786-4. doi: 10.1145/2016656.2016661.

[146] Markus Schedl, Christian Höglinger, and Peter Knees. Large-scale music exploration in hierarchically organized landscapes using prototypicality information. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 8:1–8:7. ACM, 2011. ISBN 978-1-4503-0336-1. doi: 10.1145/1991996.1992004.

[147] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6):103–108, 2017. doi: 10.9781/ijimai.2017.09.001.

[148] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. Cartograph: Unlocking spatial visualization through semantic enhancement. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 179–190. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3025171.3025233.

[149] Sylvain Senecal and Jacques Nantel. The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169, 2004. doi: 10.1016/j.jretai.2004.04.001.

[150] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992. doi: 10.1145/102377.115768.

[151] Ben Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996. doi: 10.1109/VL.1996.545307.

[152] Ben Shneiderman. *Human-Centered AI.* Oxford University Press, 2022. ISBN 9780192845290.

[153] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 830–831. ACM, 2002. ISBN 1-58113-454-1. doi: 10.1145/506443.506619.

[154] E. Isaac Sparling and Shilad Sen. Rating: how difficult is it? In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 149–156. ACM, 2011. ISBN 978-1-4503-0683-6.

[155] Anselm Strauss and Juliet Corbin. Grounded theory methodology. In Norman K. Denzin and Yvonna S. Lincoln, editors, *Handbook of qualitative research*, pages 273–285. SAGE Publications, Thousand Oaks, CA, US, 1994.

[156] Emily Sullivan, Dimitrios Bountouridis, Jaron Harambam, Shabnam Najafian, Felicia Loecherbach, Mykola Makhortykh, Domokos Kelen, Daricia Wilkinson, David Graus, and Nava Tintarev. Reading news with a purpose: Explaining user profiles for self-actualization. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP'19 Adjunct, pages 241–245. Association for Computing Machinery, 2019. ISBN 9781450367110. doi: 10.1145/3314183.3323456.

[157] Kirsten Swearingen and Rashmi Sinha. Beyond algorithms: An hci perspective on recommender systems. In *Proceedings of the SIGIR 2001 Workshop on Recommender Systems*, 2001.

[158] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Scalable collaborative filtering approaches for large recommender systems. *J. Mach. Learn. Res.*, 10: 623–656, 2009. URL `http://dl.acm.org/citation.cfm?id=1577069.1577091`.

[159] Nava Tintarev and Judith Masthoff. Explaining recommendations: Design and evaluation. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 353–382. Springer US, Boston, MA, 2015. ISBN 978-0-387-85819-7. doi: 10.1007/978-1-4899-7637-6_10.

[160] Nava Tintarev and Judith Masthoff. Beyond explaining single item recommendations. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 711–756. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_19.

[161] Nava Tintarev, Shahin Rostami, and Barry Smyth. Knowing the unknown: Visualising consumption blind-spots in recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, pages 1396–1399. ACM, 2018. ISBN 978-1-4503-5191-1. doi: 10.1145/3167132.3167419.

[162] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240, 1970. doi: 10.2307/143141.

[163] Marc Torrens and Josep-lluís Arcos. Visualizing and exploring personal music libraries. In *Proceedings of the 5th International Conference on Music Information Retrieval*, ISMIR'04, pages 421–424, 2004.

[164] Chun-Hua Tsai. An interactive and interpretable interface for diversity in recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 225–228. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3030024.3038292.

[165] Chun-Hua Tsai and Peter Brusilovsky. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 239–250. ACM, 2018. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172959.

[166] Chun-Hua Tsai and Peter Brusilovsky. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 391–396. ACM, 2019. ISBN 978-1-4503-6272-6. doi: 10.1145/3301275.3302318.

[167] Chun-Hua Tsai and Peter Brusilovsky. Exploring social recommendations with visual diversity-promoting interfaces. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(1):5:1–5:34, 2019. doi: 10.1145/3231465.

[168] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 31–40. ACM, 2007. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240630.

[169] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 109–116. ACM, 2011. ISBN 978-1-4503-0683-6. doi: 10.1145/2043932.2043955.

[170] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362. ACM, 2013. ISBN 978-1-4503-1965-2. doi: 10.1145/2449396.2449442. URL `http://doi.acm.org/10.1145/2449396.2449442`.

[171] Paolo Viappiani, Boi Faltings, and Pearl Pu. Preference-based search using example-critiquing with suggestions. *Journal of artificial intelligence Research*, 27:465–503, 2006.

[172] Jesse Vig, Shilad Sen, and John Riedl. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '09, pages 47–56. ACM, 2009. ISBN 978-1-60558-168-2. doi: 10.1145/1502650.1502661.

[173] Jesse Vig, Shilad Sen, and John Riedl. Navigating the tag genome. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 93–102. 2011. doi: 10.1145/1943403.1943418.

[174] Wesley Waldner and Julita Vassileva. A visualization interface for twitter timeline activity. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*, page 45, 2014.

[175] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

[176] Bo Xiao and Izak Benbasat. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, 31(1):137–209, 2007.

[177] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 14(1):1–101, 2020. doi: 10.1561/150000006.

[178] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matú\v s. Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4511–4515, 2010. doi: 10.1073/pnas.1000488107.

Appendix

The following article is reused from:

Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 183–191. ACM, New York, NY, USA, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394841

Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study

Thao Ngo[*]
thao.ngo@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

Johannes Kunkel[*]
johannes.kunkel@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

Jürgen Ziegler
juergen.ziegler@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

ABSTRACT

While online content is personalized to an increasing degree, e.g. using recommender systems (RS), the rationale behind personalization and how users can adjust it typically remains opaque. This was often observed to have negative effects on the user experience and perceived quality of RS. As a result, research increasingly has taken user-centric aspects such as transparency and control of a RS into account, when assessing its quality. However, we argue that too little of this research has investigated the users' perception and understanding of RS in their entirety. In this paper, we explore the users' mental models of RS. More specifically, we followed the qualitative *grounded theory* methodology and conducted 10 semi-structured face-to-face interviews with typical and regular Netflix users. During interviews participants expressed high levels of uncertainty and confusion about the RS in Netflix. Consequently, we found a broad range of different mental models. Nevertheless, we also identified a general structure underlying all of these models, consisting of four steps: data acquisition, inference of user profile, comparison of user profiles or items, and generation of recommendations. Based on our findings, we discuss implications to design more transparent, controllable, and user friendly RS in the future.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → User studies**.

KEYWORDS

Recommender Systems, Transparent AI, Mental Models, Grounded Theory, Think Aloud

ACM Reference Format:
Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3340631.3394841

1 INTRODUCTION

With the growing use of intelligent algorithms in current systems, such as recommender systems (RS), end-users find it increasingly hard to comprehend the rationale behind a certain recommendation. Thus, it is important for users to understand the relationship between user input and recommendation of RS [36]. In most cases, systems appear to users as black boxes, particularly in case of the increasingly used complex probabilistic techniques [12]. Previous research suggests that this opaqueness can lead to feelings of discomfort or even creepiness when a personalized recommendation matches a user's interest very accurately [41]. These feelings, in turn, may have negative consequences on users' trust in a RS and their intention to accept recommendations. Thus, recently, research efforts were made to increase transparency and control of a RS, e.g. through interactive explanatory interfaces [21, 42]. An important and understudied question in this context is what kind of *mental models* users form of RS. Based on in-depth knowledge about such mental models, designers of RS could make recommendations more transparent and controllable, thus mitigating the negative consequences.

Mental models can be defined as subjective knowledge representations of technological systems (e.g. computer programs) [26, 33]. Previous research indicates that users do construct mental models for RS. The soundness of these models influences satisfaction and effectiveness of interaction with the RS [9, 16]. As such, mental models focus on practical effectiveness and on making predictions about the outcome of the system. They are typically incomplete, inaccurate, and may contain areas of uncertainty [26, 33].

Due to this subjective nature of mental models, a qualitative approach seems to be most appropriate to investigate them. This approach allows us to investigate the users' unique perspectives in-depth and ask for *what* and *why* users hold certain mental models of a RS. Specifically, we chose the *Grounded Theory* (GT) methodology [5] due to its strong exploratory and data-driven nature. The participants' knowlegdge, experiences, and attitudes solely drive the data collection and analysis. Thus, the results from this methodology emerge from the data. In other words, they are *grounded* in them.

In GT, data sampling is performed purposefully, i.e. not randomly. Thus, to reveal what mental models users of RS, what assumptions these models entail, and what implications for future RS development can be derived from them, we focused on mental models of typical and regular RS users. In particular, we aim to answer four central research questions:

- RQ1: What are the mental models users hold of a RS?
- RQ2: To what extent is the RS perceived as transparent?
- RQ3: To what extent is the RS perceived as controllable?
- RQ4: What implications for RS design can be derived?

[*] Both authors contributed equally to this research.

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394841

In this study, we chose Netflix as an example because it makes extensive and apparent use of recommendations [11]. Moreover, it is one of the most popular video-on-demand services in the U.S. and Germany [8, 38]. Thus, the sample of this study most likely has developed a mental model of Netflix.

We make two main contributions with this work: (1) A theoretical contribution in form of the exploration of mental models of RS. The mental models provide in-depth insights to the user assumptions of how a RS works internally. For example, we found that all mental models followed a basic structure, comprising four steps: data acquisition, inference of user profile, comparison of user profiles or items, and generation of recommendations. (2) A practical contribution in form of discussing how our theoretical results can be applied to the development of RS. For instance, we suggest to link recommendations and user preferences more explicitly than it is done to this date.

2 BACKGROUND AND RELATED WORK

RS have become widely adopted tools to pro-actively filter online content with respect to the current user's preferences. While recommendation algorithms are able to suggest items with high precision, quality criteria that go *beyond accuracy* [15, 24] were neglected for a long time. It has been argued that user-centric aspects, such as the system's perceived transparency or the degree of control users are able to exert, constitute important facets of a system's overall perceived quality [2, 29].

2.1 Transparency and Control in RS

Typically, RS appear as *black box* to their users as it remains opaque why items are recommended and how they relate to the users' preferences [13, 35]. Increasing the transparency of a RS constitutes a prominent issue in HCI design for RS [2, 9]. It can improve perceived quality of recommendations [18], their acceptance [6, 13], and users' confidence [36]. Therefore, many researchers have called for explainable RS, i.e. the increase of system transparency through (mostly textual) explanations (e.g. [40, 42]).

Another aspect that goes *beyond accuracy* is the extent to which users can exert control over the recommendation process. Allowing users to control what is recommended to them can increase user satisfaction [32] and the perceived accuracy of predictions [28]. While many RS rely on user ratings (e.g. implicitly by recording click-through streams, or explicitly by eliciting thumb up/down ratings) [31, 37], more advanced methods for controlling recommendations have been suggested. Examples include relating preferences and recommendations more directly [1, 19], or eliciting preferences for groups instead of single items [3, 22].

Transparency and control are not independent from each other. To exert control over their recommendations effectively, users need insights into the system's reasoning—at least to a certain degree [9, 40?]. Yet, the relation between transparency and control is not trivial to investigate and may lead to counter-intuitive observations. Tsai and Brusilovsky [42], for instance, found that, besides increasing transparency, explaining recommendations can also result in a *decrease* of the perceived degree of control. According to the authors, this might be due to information overload effects entailed by the explanatory interfaces.

Such observations underline that putting transparency and control into practice may not be straightforward. In this context, we add another aspect that might be responsible for this: a discrepancy between a user's *mental model* of a system and its actual behavior.

2.2 Mental models in RS

Mental models can be defined as knowledge representations of technological systems, which are generated through interaction with the respective system [26, 33]. Rumelhart and Norman [33] used the terms of *represented* and *representing world*. The mental model represents an object or a situation of the represented world inside the cognitive representing world. This points out that mental models are *constructed*, i.e. the representing world is incomplete as it only contains those properties of the represented world that were deemed necessary. Elsewhere, Norman [26] uses a slightly different terminology to which we adhere in this paper: Based on a *target system* (i.e. the represented world) the user invents a mental model (i.e. the representing world) to simulate system behavior and make assumptions about interaction outcomes. Norman underlines that the users' mental models are incomplete, contain areas of uncertainty and possibly superstition, and focus on practical effectiveness rather than technical accuracy. In contrast to the user's mental model, the *conceptual model* represents a *more appropriate* model of the target system in terms of accuracy, consistency and completeness. They are constructed by specialists regarding the target system (e.g. the system designers).

Yet, mental models need some degree of technical correctness to let users successfully predict system behavior and thus, use it effectively. If this is not the case, misaligned mental models can result in what Norman describes as "gulfs" between user and system [27]: The *gulf of execution* occurs when a user's mental model is erroneous in terms of how a specific task can be performed with the system. The *gulf of evaluation* occurs when the actual outcome of an action with the system diverges from what the user's mental model predicted. These gulfs are well-known in usability engineering and account for many problems and misconceptions arising in HCI. One reason for the occurrence of such gulfs may lie in the transfer of a mental model from one technical system to another. To save cognitive effort, users try to re-use mental models whenever it seems feasible [26, 27].

Shneiderman and Maes concluded that one important future challenge is to make users aware of how autonomous software agents (e.g. RS) came to decisions and thus, become predictable for users [34]. Even though they did not use the term of mental models explicitly, they described them implicitly as making practical predictions about the outcome of the system is the most central utility of mental models. Surprisingly, this aspect was not further investigated in the subsequent years. To this date, the literature on mental models of RS is relatively sparse.

Only few studies have examined mental models in the context of RS so far. In an initial online survey, Ghori et al. [10] presented scenarios of different RS platforms and asked for users' knowledge and beliefs about RS. While they did not explicitly elicit mental models of RS, they concluded that users hold a "cognitive model", understand that RS track user behavior, and have rudimentary ideas of filtering mechanisms. In an exploratory approach, Kodama et al.

[14] elicited different mental models that middle school students create of the Google search engine. They found, that these models were most often wrong and conclude that concepts behind algorithmic agents should be taught better. In line with this, Kulesza et al. [16] has shown in an experiment that users who increase soundness of their mental model during usage were more efficient in controlling their recommendations. This resulted in a higher satisfaction with the outcome. To operationalize the systematic consideration of users' mental models into actual software design, Eiband et al. [9] have proposed a stage-based, iterative prototyping approach that targets at making RS more transparent through offering explanations.

3 METHOD

The research goal of this study is to investigate the users' mental models. Since the structure of mental models is inherently subjective and individual, a quantitative approach would be insufficient for this goal as this approach aims at analyzing empirical data for predetermined hypotheses. Thus, to explore unknown and highly individual mental models, we deem a qualitative approach as more appropriate.

Our qualitative study followed the *Grounded Theory* (GT) methodology [5, 39]. GT is an established and well-defined methodology from social sciences for systematic data collection and analysis. This methodology has a strong exploratory focus, i.e. no clear theory about the topic at hand is presupposed. Concepts evolve from the data during conduct of the study and hence, are *grounded in* the data. Due to the lack of predetermined hypotheses, data sampling follows the approach of *theoretical sampling* [5]. This means that sampling is performed purposefully, not randomly. Furthermore, while sampling in quantitative research is typically randomized and person-wise, in qualitative research theoretical sampling is done iteratively and concept-wise. Data are collected, coded, and analyzed simultaneously. In this way newly occurring concepts determine the sampling during the study to explore them dynamically. For this, the differences in relevant concepts (also called contrasts) are deliberately varied until no further novel observations regarding the concept are made. Then, the state of so-called *theoretical saturation* [5, 25] is achieved. In this case, either another concept is explored or the study is concluded if the pursued theory is already well-developed.

In our study, the *theory* of GT are the different mental models users hold of a RS. We deliberately focused Netflix as an example for RS, since it 1) is well-known for its extensive and apparent use of recommendations [11], and 2) is wide-spread, increasing the likelihood for us to sample a broad variety of contrast in our concepts. In other words, this allows us to study different variations of one concept. As instruments we applied individual semi-structured face-to-face interviews, which we combined with a *Think Aloud* task and a drawing task to capture different facets of each mental model as broadly as possible. Following the approach of theoretical sampling, we deliberately recruited participants of whom we had information about their background and who fit to the current concept under consideration (e.g. the level of technical knowledge). Throughout the entire study, we only sampled participants with advanced Netflix experience (frequent use for at least one year), as we aimed to focus on the typical Netflix user. All in all, we recruited ten participants (six female) with an age range between 19 and 31 ($M = 24.70$, $SD = 4.57$). Hence, our sample represented the typical Netflix user group well [7]. The interviews were conducted in July and August 2019.

For our analysis, each interview was transcribed in a timely manner using easytranscript 2.50 and analyzed with MAXQDA 18.2.3. The transcribed interview of each participant was first coded by two independent raters. Subsequently, the two raters discussed and analyzed each interview jointly. During analysis, various analytic tools and mental strategies were used, including *microanalysis* of the data through open *line-by-line coding*, *constant comparison* and *axial coding* to summarize the open codings to categories, and *selective coding* to infer the mental model of Netflix for each participant. To ensure that codes and resulting categories emerge from the data, throughout the whole process *in-vivo codings* (i.e. verbatim codes from participants' statements) played a central role. In order to record impressions, evolving theoretical concepts and the relationships among them, raters made extensive use of memos, which constitutes a substantial aspect of the GT method.

This iterative process led to 10 distinct categories such as *evaluation strategy for items*, which pertains to how participants asses the quality of items (e.g. *content-based* vs. *non-content-based*), *search strategy for items*, which describes how participants decide whether to consume an item or not (e.g. through *internal* or *external* information acquisition), and *general model of RS*, which we focus within the scope of this work and which emerged from our data presented in Section 4.

3.1 Preparation

The study was approved by the local ethics committee of the University of Duisburg-Essen in Germany. Participants consented to the interview and audio recording. All personally identifiable information was anonymized.

3.1.1 Interview guide. We developed an interview guide with an interview duration of roughly one hour. All interview questions in the guide were open-ended. The interview started with a brief introduction of the interviewer and a short description of the purpose and motivation of the interview. It was emphasized that the study is concerned with the recommendation component of Netflix and that the main interest of the study lies in the exploration of the participants' experience with the personalized content of Netflix. The participants were then asked about their experience with Netflix: How often do they use Netflix? Since when do they use it? Which experience do they have with the recommendations in Netflix? Which parts of Netflix are subject to personalization? How confident do they feel with personalization in general?

The interview proceeded with the Think Aloud task: Participants were instructed to use their own Netflix account to find a comedy movie to watch in the evening that was in line with their preferences. After that, a hypothetical scenario was introduced. Participants had to imagine that the movie was not as good as expected. Thus, they should try to express negative feedback to Netflix for that movie. The purpose of this task was for the participants to reflect on the options to express their preferences to Netflix. Then, participants were asked about the functioning and data processing of Netflix: How does a RS like Netflix work? Which data are used by the system? What happens to the data in order to generate recommendations? Do they know about the thumb function of Netflix? What does it trigger?

Finally, participants received a sheet of paper and were asked to draw their very own image of Netflix. Participants were informed that they could perform this task freely, without any limitation. They were prompted to explain their drawing.

At the end of the interview, participants were debriefed and offered to ask questions and give general feedback on the interview. No incentives were given for participation, besides a certificate of taking part in the study[1].

3.2 Data sampling

Our data sampling was fundamentally influenced by the data-driven approach of GT and of theoretical sampling. Accordingly, we sampled participants not at random, but based on what concepts we decided to explore next. In general our data acquisition was organized in three phases with different foci. In the following section, we elaborate on our sampling decisions for this study.

3.2.1 First sampling phase: Typical Netflix users. As recommended by Corbin and Strauss [5], we first focused on a typical sample of the target population. According to Dahlgreen [7] 57% of Netflix users are female and roughly 50% of all Netflix users are between 18 and 34 years old. Our initial sample (P1, P2, P3, P4) were females with an age of 21, 24, 27, and 24. Thus, we were able to recruit a sample within the age range of typical Netflix users.

In this first sampling phase, we found the concepts of *centrality of self* and *item-based recommendation.* The first concept was especially salient in the interview of P3, while the item-based recommendation was most apparent in the drawing of P2. We found these concepts mostly through *comparison.* During further *axial coding*, we observed that all participants held rather *technical* mental models (see Section 4.3), i.e. they are close to the functioning of algorithms or procedures. This became mostly apparent through *microanalysis*, which revealed that P1, P3, and P4 generally used many technical terms (e.g. "*ip address*" (P1, P3), "*database*" (P1, P3, P4), and "*dynamic query*" (P3)).

Following the *flip-flop technique* [5], we turned this concept "upside down", asking ourselves questions such as: "*How are the mental models in case of lower technical knowledge?*", and "*How are the mental models in case of higher technical knowledge?*" In order to investigate these questions, we decided to sample low and high extremes on the dimension of technical knowledge next.

3.2.2 Second sampling phase: Low/high technical knowledge. Next, we purposefully sampled P5 and P6. Both participants were male and aged 31 and 30, respectively. While P5 had a very low technical background regarding RS (he held a bachelor's degree in arts and was currently unemployed), P6 had a high technical knowledge as he worked in computer science research and was currently engaged with decision support systems.

In this second sampling phase, we found mental models which differed in a *metaphorical* and *technical dimension* (see Section 4.3). P5 clearly held a metaphorical mental model: He drew Netflix as a monster serving recommendations with many arms (see Figure 2c). In contrast to this, P6 had a technical idea of Netflix.

In addition, P6 mentioned that he used the explicit rating function (thumbs up/down) occasionally to steer his Netflix account towards better recommendations. We found this aspect quite striking as we did not elicit any responses on what influence the explicit rating function may have until this point of the study. Rather tentative, we assumed a connection between usage of explicit ratings and decided to restrict our next sample to participants using explicit ratings frequently.

3.2.3 Third sampling phase: Use of explicit ratings. During this third sampling phase, we conducted interviews with participants P7, P8, P9, and P10 aged 19, 19, 22, and 30. P7 and P10 were male, while P8 and P9 were female. All declared using the thumbs function in Netflix frequently. During analysis of this sample, we observed the counterpart of *item-based recommendation*, namely *user-based recommendation* (see Section 4.2).

After analyzing the data, we found that the main concepts, which we found before, did not show further variation in these observations. Thus we deemed them as theoretically saturated. Especially, with regards to our main aspects of transparency and control, we did not see new insights in the interviews. Thus, we ended our data sampling at this point.

4 RESULTS

Overall, one general structure of users' mental models of RS emerged from our collected data. All participants followed the same pattern and divided the functioning of RS into four separate steps: (1) *data acquisition*, (2) *inference of a virtual user profile*, (3) *comparison of user profiles or items*, and (4) *generation of recommendations* (see Figure 1).

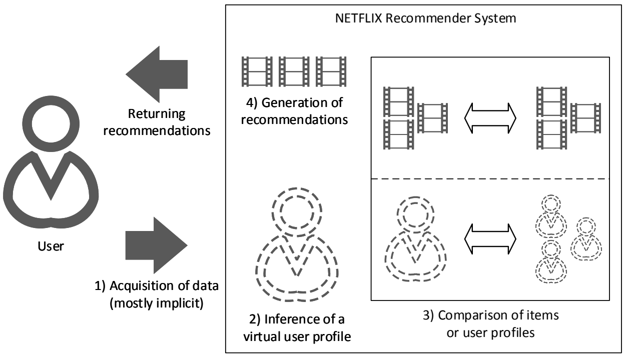

NETFLIX Recommender System
Returning recommendations
4) Generation of recommendations
User
1) Acquisition of data (mostly implicit)
2) Inference of a virtual user profile
3) Comparison of items or user profiles

Figure 1: Basic mental model found in all participants.

Regarding the acquisition of data in step (1), our analysis revealed that participants considered user characteristics, such as location, gender, and age, as well as user interaction behavior as relevant for Netflix. For the latter, we were able to form two categories: *implicit* user behavior, such as watching a movie, and *explicit* user behavior, such as pressing the thumbs-up button. From these data, participants assumed that Netflix derives a virtual user profile in step (2). This profile may contain *latent* item characteristics, which

[1] This was requested by 3 of the 10 participants since they needed to participate in empirical studies as part of their study program.

are not visible to the user. For example, P5 speculated: "*As far as I know, there are a lot of subcategories in the background which a user does not see on the interface.*" In step (3), participants assumed that comparisons between items or user profiles were made. These two general directions adhered to the concepts of *user-* and *item-based recommending* (see Section 4.2). Finally, step (4) corresponds to the actual selection of personal recommendations. Here, participants assumed that all process data cumulated into one recommendation. This assumption is for instance depicted in the drawing of P3 (see Figure 2b).

Regarding details of how the four steps are performed, participants made diverse assumptions. Nonetheless, across all interviews, participants expressed confusion and uncertainty when asked about the inner working of Netflix: "*I don't know which data they have of me.*" (P3), or "*It's a black box. I don't know how they do it. Maybe I should know it.*" (P7). Many of them also rejected the recommendations provided by Netflix, as P2 stated:

> "*Some [recommended] movies I find interesting, but there are also many things, I am not interested in. I feel that my preferences don't play a role, instead it's just [the movies] which people are currently talking about.*"

Furthermore, based on participants' drawings and statements, we derived the concept of "centrality of self" from our data as well as two dimensions that characterize the identified mental models. They are reported in the following sections.

4.1 Centrality of self

Some participants clearly viewed their own self as central component in their Netflix experience (P3, P5, and P6). This became particularly apparent in the drawing of P3 (see Figure 2b), as she confidently started the drawing task with the role of herself ("*Ok, I am still a little overburdened by what to begin with. Ok, in any case, first of all we need myself: the Netflix user.*"). Then, the entire drawing evolved around this central self. While many other participants used a content-based approach of explaining how recommendations are generated (see Section 4.2), content aspects of any kind were entirely absent in the drawing of P3. Instead, in different parts of the sketched functionality, users played a central role, for instance, when recommendations are generated based on what was watched before (box with dashed line in the southwest of Figure 2b). The *centrality of self* together with the importance of users per se and their social interrelation, was emphasized by the participant's estimation of an existing internal connection between Netflix and Facebook (northern arc in Figure 2b). P3 assumed that the history of what her friends watched in the past was also taken into account when recommendations for herself are generated, and vice versa. Note, that this drawing clearly depicted three of the four general steps discussed above: data is acquired (watched items and Facebook data), similarity is calculated between users and items (inside the box with the dashed line), and recommendations are generated (arrows inside the box on the right).

The concept *centrality of self* can also be found throughout the interview of P3. She, for instance, mentioned that her recommendations are sometimes inaccurate. This results in long searching sessions, which she described as tedious and confusing. Yet, the reason for this lack of decisiveness was sought at her own side:

> "*Because it takes an extreme amount of time to search, but I would not necessarily burden this on Netflix but on myself, since I am never satisfied with my choice.*"

Note, however, that when being asked, P3 did not assess herself as a person who has a hard time to decide in general ("*at the supermarket [. . .] I am very determined.*"). Even though, this person did not ascribe the problem of long searching sessions in Netflix to the RS, she fancied the idea of having better explanations for her recommendations. In particular, P3 formulated a wish to know more about the relation of recommendations and her own preferences. As a consequence, the category of "similar to . . . " recommendations was perceived as helpful, yet also as arbitrary. When confronted with the idea to be able to control to which preferences recommendations are generated, P3 expressed a strong affection for such a feature[2].

Other participants did not see the self as central as P3 but elaborated on the role of the self implicitly throughout the interview. P1, for instance, showed some aspects of *centrality of self*, when she was asked to clarify the difference between implicit and explicit ratings. She underlined that her explicit ratings have higher impact compared to her implicit interaction data because she used the thumb function seldom. In the same answer during the interview, P1 took over the role of Netflix talking about herself: "*Ok, now she clicked on something [i.e. rated an item], so we will give her more of that.*" Even though rather shallow, the *self* as a concept was mentioned in both statements. It constituted a counterpart to Netflix as a system making assumptions about the user. A similar effect could be observed in answers of P10. He emphasized his own responsibility for the influence on recommendations ("*If I dislike Adam Sandler but all the time [. . .] watch movies starring him, I do not have to wonder when a Adam Sandler comes out [of the RS].*").

4.2 User- vs. item-based recommendations

When participants were asked about the rationale they assumed behind items being recommended, we observed two major directions. While some participants assumed recommendations being generated with respect to similarity between items (P5, P8, P10), others followed a user-based approach (P1, P9). As P9 put it:

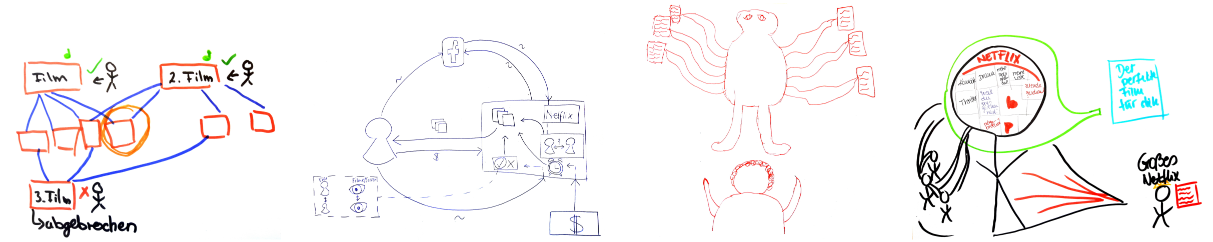
> "*[Recommendations base on] other users: what other users frequently watched, or gave a good rating for.*"

Strikingly, this style resembled closely the explanation model used by *Amazon* ("*Users who bought . . . also bought . . .*"). This resemblance was also explicitly mentioned by P9:

> "*I could imagine that it is like e.g. at Amazon. There it is also written that users bought something together.*"

P9 adhered to this form of thinking in her drawing as well (Figure 2d). Here, the entire process of generating recommendations was envisioned as inherently social. It depicted a crowd of users at the bottom left, which was connected to the personal Netflix agent (large stick figure in the middle). Through this connection the agent selects a movie as recommendation. Even Netflix in general was depicted as person, which instructs and overviews the entire process (at the bottom right of the drawing). When examining level

[2] Note, that such a function actually exists (next to the details for a movie or TV show). P3 also knew this function but, nonetheless, wished it to be more visible and that the "similar to . . . " category on the front page was replaced by an interactive version.



(a) Drawing of *P2.* A user watches three movies, from which two are liked. The recommendation (orange circle) is similar to the liked ones.

(b) Drawing of *P3.* The *centrality of self* is highly salient as the entire recommendation process evolves around this person's self.

(c) Drawing of *P5.* Netflix as a "tentacle monster", which can handle a huge range of recommendations with its many arms.

(d) Drawing of *P9.* The entire process of recommending is perceived as inherently social. This perspective highly emphasizes the role of users in Netflix.

Figure 2: Drawings of participants asked to illustrate their mental model of the inner workings of Netflix

of technical knowledge, P9 mentioned that she "*studies in that area*" and did have some knowledge about "*technical topics, AI, and such*".

Such social assumptions were juxtaposed by a model of Netflix that was based on content features of items. Examples for features being utilized for deriving similarity between items were "*actors*" (P8), "*buzz words*" (P10), and other content data such as "*movies set in the same time*" (P3). Another aspect for item-based comparison that was frequently mentioned were latent categories. Such categories were supposed to be only used "*in the background*" (P5) and had a finer granularity:

> "*there is not just action but also Asian action, German, and English... such things [. . .] for depicting more accurate [recommendations].*" (P6)

Apparently, this assumption originated from the RS in *Spotify* as P6 further explained: "*Once I saw a list somewhere containing Spotify genres. [. . .] They have somewhat over* 400 *genres*".

However, user- and item-based styles were not fully mutually exclusive. P2, P3, P4, and P6 showed aspects of both dimensions. P3, for instance, assumed a hybrid algorithm, which combines items watched by similar users and items that have a similar genre to the recently watched ones. Over all different styles, frequent use of verbs like *thinking*, *guessing*, and *believing* underlined the uncertainty about the inner workings of Netflix. The same applied to the *matching score*, which was shown for recommendations at Netflix. All participants agreed on being uncertain regarding what is actually *matched* when talking about the depicted score ("91 % *match – whatever that means.*" (P6)).

4.3 Technical vs. metaphorical

We found that the mental models can be characterized as either technical or metaphorical. Technical models were expressed by six participants (P2, P3, P4, P6, P7, P8). They used process diagrams and data flows to explain how Netflix arrives at its recommendations, which indicated a procedural understanding. For instance, P8 described:

> "*I am thinking about which data Netflix takes from me or already hold of me. [. . .] From this they know, what I like to watch. What else? Actors, producers... they take this from the movies I watched. Then, [Netflix] takes a look at the match and searches for [recommendations].*"

In the technical models, the general four steps were often made explicit by the participants. For example, P2 explained the steps *data acquisition*, *inference of user profile*, and *comparison of user profiles*:

> "*Probably everything is saved and collected for each user. And then, they compare users with similar profiles, in terms of the movies, to see whether these users have similar interests. Then, perhaps, one similar user has rated a movie positively the other one has not yet watched. Interviewer: And this is how they arrive at recommendations? P2: Yes, for instance.*"

The clear understanding of P2 was underlined by her drawing (see Figure 2a), which depicted the same process of recommending from an item-based perspective.

A different standpoint was taken by four participants (P1, P5, P9, P10), who used a metaphorical description of Netflix and thus, drew characters to illustrate how the RS works. P1, for instance, used a metaphor of a house:

> "*A huge complex house in which all the data and databases are somehow saved. [. . .] Of course, there are employees, but I think everything works with algorithms.*"

P1 focused on Netflix as a whole entity and in a more literal way than other participants. She expressed the four basic steps in the interview, however, for the drawing task, she chose the depiction of a house. This could be seen as a simplification of Netflix and a tangible understanding of Netflix which was based on the Netflix corporation building.

Additionally, some metaphorical mental models clearly entailed the participant's attitudes towards Netflix. P5 compared Netflix to a tentacle monster (see Figure 2c):

> "*It has all its tentacles and at each tentacle it offers its products, the movies it has. It's like a kraken monster. It has a huge range of offers, hence so many tentacles so that there is something for everybody.*"

This description expressed a negative view on Netflix. When browsing through the catalog of movies to find a matching one during the Think Aloud task, this participant expressed feelings of being lost and confused:

> "*Most things here mean nothing to me. These all are arbitrary and random images that do not catch me so that I think I don't want to look into [the details of the movie]. No. [. . .] everything seems absolutely random.*"

Such negative feelings were also quoted by other participants. Some, for instance, pointed out that personalization of Netflix engenders a loss of diversity in their movie consumption and therefore pose a risk of becoming trapped in a "filter bubble" (e.g. P1, P7, P9).

Apart from that, nearly all participants lacked trust in Netflix. Especially they doubted the system's integrity assuming that recommendations were biased towards in-house productions or third-parties, "*I have the feeling that in-house productions are mostly advertised and this is not necessarily good.*" (P6). When being asked about who influences the movie recommendations, P3 mentioned third-parties: "*The producers of the movies. [. . .] And perhaps record companies of movie soundtracks?! I don't know.*" Hence, at least some participants were aware of the economic interest of Netflix and third-parties. However, P9 justified their influence on the personalization process: "*I think it is good how it is because they do not exaggerate it and draw attention to it. It is also their production. [. . .] Therefore, it is good and also their right to advertise for themselves.*"

5 DISCUSSION

Regarding our first research question (" *What are the mental models users hold of the RS?*"), we found very diverse mental models, which, nonetheless, all adhered to a very basic structure—even among those participants with little technical knowledge. This structure consists of four steps: *data acquisition*, *inference of a user profile*, *comparison of items or users*, and *generation of recommendations* (see Figure 1). As this basic structure was held by all participants, we suspect that this structure might be prevalent in many typical and regular Netflix users. Our results extend the findings by Ghori et al. [10] substantially through the identification of this general model.

The subsequent sections are organized regarding our other research questions thus, asking for transparency and control in the identified mental models, and finally for possible implications for RS design.

5.1 As how transparent is Netflix perceived?

Across the four general steps, participants made various causal assumptions of how recommendations are derived and how their behavior as user affects them. We observed many discrepancies regarding these assumptions during single interviews, and especially between participants' drawings and their explanations throughout the rest of the interview. Assumptions were highly speculative and led to confusion—even superstition. This resulted in an effect we term *mystification* of the underlying RS: Participants invented various suppositions about the capabilities of the system, although they might be entirely unjustified and lack realistic evidence. One example illustrating this is P3's assumption that she receives recommendations, based on what her friends liked on Facebook. Thus, regarding RQ2, we conclude that users did not perceive the RS of Netflix as very transparent. We note that this was also not mitigated by the experience in using a RS, since we observed this in spite of the rather advanced experience with Netflix all participants had.

As a consequence of this lack of transparency, users encountered a *gulf of evaluation* (i.e. users did not understand what their recommendations were based upon) and were thus not able to exploit the full potential a personalized RS bears. We also found that mystified beliefs may harm the reputation of Netflix, which is shown by metaphorical mental models entailing negative attitudes towards the RS (e.g. P10 cynically drew Netflix as evil hungry black box eating user data and "pooping" recommendations). Reasons for this *mystification* and *gulf of evaluation* can be found in the dimensions we identified as concepts.

Participants, showing the *centrality of self* (Section 4.1), were clearly aware of the role of their own self, which we assume to be a general stance when encountering the surrounding world. Not surprisingly, this was also applied to the interpretation of recommendations. The users who expressed the centrality of self, wished to be more informed about which information about them is responsible for the recommendations. Participants were not able to understand this causality, which also resulted in a *gulf of evaluation* and, consequently, in a demand for a higher transparency regarding the influence of user preferences on recommendations.

In line with Norman [26], we observed that many of our participants tried to transfer their mental model of Amazon to Netflix. The RS of Amazon provides users with textual explanations for recommendations (i.e. products that were bought together with the currently inspected one). These explanations follow the algorithmic mechanism of item-based collaborative filtering, which we also found in the concept of *item-based recommending* (Section 4.2). While the prevalence of such algorithmic methods in the users' mental models, might be beneficial in some special cases (i.e. when source and target RS are algorithmically very similar), we assume such situations to be rather unlikely in practice. Our observations, for instance, show that the transfer of the mental model of Amazon to Netflix lead to false assumptions and misunderstandings. We ascribe this mainly to the different forms of how recommendations are presented. While Amazon follows an item-based approach, showing recommendations right next to the textual specification of single products, Netflix mainly presents recommendations in accordance to the entire user profile (i.e. "*top picks for you*"). Consequently, participants were highly unsure about how the list of recommendations was constructed.

5.2 As how controllable is Netflix perceived?

In our third research question, we asked ourselves to what degree the RS of Netflix is perceived as controllable by its users. As mentioned (e.g. in [9, 40?]), transparency and control are interdependent. We observed the same in our study: The lack of transparency led to a *gulf of execution* (i.e. participants were unable to figure out what interaction possibilities they had). Consequently, they also found it unclear how to steer the RS towards recommendations fitting their needs more adequately.

One reason we deem responsible, is again the transfer of mental models from Amazon to Netflix, mainly because the rather simplistic style of explanations provided by Amazon does not provide any direct entry points for interaction: Users might perceive that they cannot influence what "users who bought, also bought". As a consequence, many participants experienced no or little control over their recommendations, although they were aware of explicit interaction options (e.g. in form of expressing a like for a movie).

To make interaction with RS less confusing, more transparent, and controllable, we argue that mental models of RS need to be aligned with the conceptual model, which represents the actual algorithmic functioning. In other words, users need to be educated about how recommendations are derived and what possibilities for interactively controlling them they have. In such a way educated users understand recommendations and their causality better, are able to use the system more effectively, and thus, are more satisfied with it and the resulting recommendations.

5.3 Implications for RS development

Considering *RQ4* ("*What implications for RS design can be derived?*"), we derive four guidelines for the development of RS. While we are aware that these are based on one particular RS, we are confident that they pose valuable anchor points for general RS design.

5.3.1 Link components to existing mental models. To reduce confusion and cognitive complexity, RS developers might rely on our identified basic mental model (Figure 1) to be already present. In particular, we encourage developers of RS to increase transparency by relating components of their system to one or more of the model's four steps. This implies that it might not be necessary to explain each single step of the inner working of RS to users in detail.

5.3.2 Align UI components with recommendation algorithm. We suggest to align explanatory and interactive components with the underlying algorithmic pattern of recommending more precisely and explicitly. Here, especially item- and user-based recommending should be distinguished. Our results indicate that both pertain to diverse mental models and that they were transferred between RS, which caused many false expectations about system behavior. In this sense, prevalent mental models might need to be corrected regarding the system's actual functioning.

5.3.3 Heed the centrality of self. RS developers should emphasize the impact of the users' current preference profile on recommended items. We particularly suggest to link content features between consumed and recommended items, since we observed that the content of items is a paramount expected aspect in the process of recommending (see Section 4.2). This does not mean that the RS has to solely rely on content-based filtering though. There is some research on how to combine collaborative filtering with content data [20, 21, 23], which could be used to make systems based on collaborative filtering more transparent using the content of items. When communicating the relation of preferences and recommendations adequately, it can also be used to exert control over recommendations (see, e.g., [1, 19]).

5.3.4 Enlighten the mystification. A central challenge of making RS more transparent and controllable is to overcome the *mystification* of RS. While this is implicitly also addressed by the guidelines above, we observed that mystification was especially a result of metaphorical mental models. Hence we suggest to introduce standardized and accordingly aligned metaphors that correct or replace existing ones. This could, for instance, be achieved by personifying the RS, e.g. by depicting an anthropomorphic avatar. However, while the depiction of such avatars and the social presence they emit, were observed to improve trust and adoption of recommendations [17, 30], negative emotions may be triggered, e.g. due to *uncanny valley* effects [4]. Thus we deem the design of feasible metaphors for RS as distinctively challenging and emphasize that it requires further research in this topic.

5.4 Limitations

Despite the small size of $N = 10$, we consider our identified concepts as theoretically saturated because we noticed that the concepts of the mental models were very well developed early in the recruitment process. The main limitation of our work is the focus on a very specific sample of *one* single platform, namely regular and experienced Netflix users which most likely has contributed to the early theoretical saturation. Finally, due to the qualitative nature of this study, we cannot make assumptions about the prevalence of the identified mental models.

6 CONCLUSIONS AND FUTURE WORK

Applying a qualitative approach, we found a variety of mental models. Our participants expressed high degrees of uncertainty and confusion about the inner working of Netflix. Nonetheless, we elicited a general structure that all of these models adhered to which can be used for RS development in practice. Furthermore, the concepts of *centrality of the self* and *item- and user-based recommending* can serve as entry points for the design of transparent and controllable RS. Hence, this work contributes not only to the exploration of users' mental models of RS, but also provides insights for RS development in practice.

In future work, we plan to validate our findings through quantitative research. Especially, the general structure represents a solid baseline for hypotheses and confirmatory studies on a large user basis. Here, it might also be interesting to investigate a more diverse user group which differ in the frequency of use and experience with RS. We stress out that it is worthwhile to investigate other RS platforms as our study focused on one single platform. Finally, the aspect of transfer of mental models was a striking result of our study. Transfer of mental models can be important for RS developers as they could rely on this to build the RS. To further investigate the transfer of mental models, we suggest to conduct comparative studies with several examples of RS.

7 ACKNOWLEDGEMENT

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under Grant No. GRK 2167, Research Training Group "User-Centred Social Media".

REFERENCES

[1] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 35–42. https://doi.org/10.1145/2365952.2365964

[2] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for Recommender Systems: The Past, the Present and the Future. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 123–126. https://doi.org/10.1145/2959100.2959158

[3] Shuo Chang, F. Maxwell Harper, and Loren Terveen. 2015. Using Groups of Items for Preference Elicitation in Recommender Systems. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*

(CSCW '15). ACM, New York, NY, USA, 1258–1269. https://doi.org/10.1145/2675133.2675210

[4] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (March 2019), 539–548. https://doi.org/10.1016/j.future.2018.01.055

[5] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Thousand Oaks, California. https://doi.org/10.4135/9781452230153

[6] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (Aug. 2008), 455. https://doi.org/10.1007/s11257-008-9051-3

[7] Will Dahlgreen. 2016. *Streaming wars: the actors Netflix and Amazon customers want to see.* Retrieved January, 15, 2020 from https://yougov.co.uk/topics/politics/articles-reports/2016/01/14/streaming-wars-actors-netflix-and-amazon-customers

[8] Deloitte. 2017. *Welchen Video-on-Demand-Anbieter nutzen Sie?* Retrieved April, 24, 2020 from https://de.statista.com/statistik/daten/studie/443820/umfrage/genutzte-video-on-demand-anbieter-in-deutschland/

[9] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961

[10] Muheeb Faizan Ghori, Arman Dehpanah, Jonathan Gemmell, Hamed Qahri-Saremi, and Bamshad Mobasher. 2019. Does the User Have A Theory of the Recommender? A Pilot Study. In *Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. CEUR-WS.org, 77–85.

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

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5, Article Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[13] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, 241–250. https://doi.org/10.1145/358916.358995

[14] Christie Kodama, Beth St. Jean, Mega Subramaniam, and Natalie Greene Taylor. 2017. There's a creepy guy on the other end at Google!: engaging middle school students in a drawing activity to elicit their mental models of Google. *Information Retrieval Journal* 20, 5 (Oct. 2017), 403–432. https://doi.org/10.1007/s10791-017-9306-x

[15] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (March 2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[16] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[17] Johannes Kunkel, Tim Donkers, Catalin-Mihai Barbu, and Jürgen Ziegler. 2018. Trust-Related Effects of Expertise and Similarity Cues in Human-Generated Recommendations. In *Companion Proceedings of the 23rd International on Intelligent User Interfaces: 2nd Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE)*. http://ceur-ws.org/Vol-2068/humanize5.pdf

[18] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[19] Johannes Kunkel, Benedikt Loepp, and 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

[20] Johannes Kunkel, Benedikt Loepp, and 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

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

[22] Benedikt Loepp, Tim Hussein, and 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

[23] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163

[24] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659

[25] Janice M. Morse. 2015. "Data Were Saturated...". *Qualitative Health Research* 25, 5 (2015), 587–588. https://doi.org/10.1177/1049732315576699

[26] Donald A. Norman. 1983. Some Observations on Mental Models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, New York, NY, USA, 7–14.

[27] Donald A. Norman. 1988. *The design of everyday things.* Basic Books, Inc., New York, NY, USA.

[28] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: Visual Interactive Recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1085–1088. https://doi.org/10.1145/1357054.1357222

[29] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[30] Lingyun Qiu and Izak Benbasat. 2009. Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems* 25, 4 (Dec. 2009), 145–182. https://doi.org/10.2753/MIS0742-1222250405

[31] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02)*. ACM, New York, NY, USA, 127–134.

[32] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 520:1–520:8. https://doi.org/10.1145/3290605.3300750

[33] David E. Rumelhart and Donald A. Norman. 1983. Representation in Memory.

[34] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *interactions* 4, 6 (Nov. 1997), 42–61. https://doi.org/10.1145/267505.267514

[35] Itamar Simonson. 2005. Determinants of Customers' Responses to Customized Offers: Conceptual Framework and Research Propositions. *Journal of Marketing* 69, 1 (Jan. 2005), 32–45. https://doi.org/10.1509/jmkg.69.1.32.55512

[36] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, USA, 830–831. https://doi.org/10.1145/506443.506619

[37] E. Isaac Sparling and Shilad Sen. 2011. Rating: how difficult is it?. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 149–156.

[38] Statista.com. 2018. *Most popular video streaming services in the United States as of July 2018, by monthly average users.* Retrieved January, 15, 2020 from https://www.statista.com/statistics/910875/us-most-popular-video-streaming-services-by-monthly-average-users/s

[39] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. In *Handbook of qualitative research*, Norman K. Denzin and Yvonna S. Lincoln (Eds.). SAGE Publications, Thousand Oaks, CA, USA, 273–285.

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

[41] Helma Torkamaan, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. How Can They Know That? A Study of Factors Affecting the Creepiness of Recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, New York, NY, USA, 423–427. https://doi.org/10.1145/3298689.3346982

[42] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining Recommendations in an Interactive Hybrid Social Recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 391–396. https://doi.org/10.1145/3301275.3302318

Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach

Johannes Kunkel(✉), Thao Ngo, Jürgen Ziegler, and Nicole Krämer

University of Duisburg-Essen, 47057 Duisburg, Germany
{johannes.kunkel,thao.ngo,juergen.ziegler,
nicole.kraemer}@uni-duisburg-essen.de

Abstract. How users interact with an intelligent system is determined by their subjective *mental model* of the system's inner working. In this paper, we present a novel method based on card sorting to identify such mental models of recommender systems quantitatively. Using this method, we conducted an online study ($N=170$). Applying hierarchical clustering to the results revealed distinct user groups and their respective mental models. Independent of the recommender system used, some participants held a strict procedural-based, others a concept-based mental model. Additionally, mental models can be characterized as either technical or humanized. While procedural-based mental models were positively related to transparency perception, humanized models might influence the perception of system trust. Based on these findings, we derive three implications for the consideration of user-specific mental models in the design of transparent intelligent systems.

Keywords: Mental models · Transparency · Recommender systems · Card sorting · Hierarchical clustering

1 Introduction

Mental models of intelligent systems are subjective, typically incomplete and flawed understandings of the system's inner working [38,45]. They are shaped through system interaction [38]. Studying mental models can, thus, explain how users perceive a system and how they interact with it, e.g. by identifying superstitions or misconceptions. This is also a crucial prerequisite to explain elements of intelligent systems better and to increase their transparency [13,54].

To investigate subjective mental models, research has focused thus far on qualitative approaches that characterize single mental models in greater detail using small samples (typically smaller than $N=20$; e.g. [13,34,36]). While such

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-85610-6_23) contains supplementary material, which is available to authorized users.

Published by Springer Nature Switzerland AG 2021
C. Ardito et al. (Eds.): INTERACT 2021, LNCS 12935, pp. 383–404, 2021.
https://doi.org/10.1007/978-3-030-85610-6_23

qualitative studies describe single, overarching models and are valuable for a general comprehension of what is included in such a model, they struggle to capture their full *diversity* and lack the ability to reliably identify and systematically compare different mental models that might coexist in large samples. Such comparisons, most importantly, offer systematic insights into relationships between specific mental models and user-centered aspects. We argue that this needs to be addressed through a quantitative approach.

In fact, quantitative methods might reveal individual and reappearing structures of mental models in *large samples* and *across* different systems. Hence, they could allow comparisons of diverse mental models among individuals and groups. Studying mental models quantitatively might also lead to practical implications for the design of user-friendly interfaces. Specific themes and visual perspectives could be designed for certain user groups, commonalities among models could foster general design of transparent systems. However, the application of quantitative methods still poses a serious challenge.

In this work, we aim to close this gap and explore the users' mental models of intelligent systems quantitatively. For this, we applied a novel card sorting setting which captured the entire processing chain of an intelligent system. The card sorting setting provided typical functional steps of intelligent systems (e.g. data acquisition) for users to reconstruct their mental model. The method allows us (1) to identify user groups and characterize their mental models, and (2) to explore the relationship of these user groups and mental models with system perceptions (e.g. transparency).

We applied this novel card sorting setting in the domain of recommender systems (RS) as RS are a mainstay in today's online environment. Furthermore, their decisions are often perceived as subjective which are met with more distrust by users than other systems that make more objective decisions (e.g. route planners) [6]. We asked RS users of a broad sample ($N = 170$) to sort different actions according to how they think the RS works internally. Hence, we aim at answering the following research questions:

RQ1: Which different mental models do users hold across RS?
RQ2: How do these mental models relate to the perception of RS?
RQ3: Based on these findings, which implications can we derive for the design of transparent intelligent systems?

With this work, we contribute to the advancement of research on users' assumptions and knowledge about intelligent systems in three ways: (1) We captured the entire processing chain of an intelligent system (i.e. RS) in a detailed way through a novel card sorting setting. (2) This allowed us to demonstrate and uncover which mental models are prevalent in a broad sample of RS users, thus forming a baseline for future research in this domain. (3) We derive practical implications for system designers regarding how the knowledge of such mental models can be leveraged to increase user-centric qualities of intelligent systems, such as transparency and trustworthiness.

2 Background and Related Work

Recommender systems typically appear as *black box* to users, i.e. their internal reasoning and functioning remain hidden. This can affect users negatively, e.g. it can cause feelings of creepiness towards recommendations [47]. Furthermore, users may distrust algorithmic decision making and reject its results [12,40]. This *algorithm aversion* seems to pertain especially to situations of subjective compared to objective decision making [6]. As a result, RS that recommend subjective items (e.g. music or movies) are more affected by distrust than objective systems (e.g. route planners that give directions) [6]. To tackle this potential distrust in subjective decision support systems, transparency appears to be a central factor. Studies indicate that transparency can increase the users' trust in and satisfaction with a system [27,49] and recommendation acceptance [11,20].

A clear understanding of the users' knowledge and interpretation of the system's functioning is a key prerequisite for determining how to improve transparency and which parts of the system to focus on [13,55]. A holistic depiction of such knowledge can be conceptualized as *mental models* [38,39].

2.1 Mental Models of Intelligent Systems

Mental models (closely related to *folk theories*[1]) can be defined as cognitive knowledge representations of technological systems that serve users to cognitively simulate system behavior and predict its outcomes [45]. They are subjective in nature, and thus, may be parsimonious and flawed [38]. Mental models are developed through system interaction, especially when confronted with anomalies and unexpected behavior [18]. In other words, mental models represent *what* users know about a system and determine *how* they interact with it.

A field study by Tullio et al. [50] demonstrated that this also holds for intelligent systems. They found that, without prior knowledge about the system, users showed a basic understanding of machine learning methods when confronted with an intelligent agent. In particular users' mental models included decision trees and statistic predictions based on "patterns" and "averages".

Other research has highlighted the impact of mental models on the users' task performance. For example, in a qualitative study Muramatsu and Pratt [34] showed that flaws in mental models of search engines may cause confusion regarding the interpretation of search results. Despite the familiarity and daily use of search engines, many participants did not fully understand how search queries are processed. This is supported by a study of Kulesza et al. [26] which showed that improved soundness of mental models was positively related to the effectiveness of interaction with the system.

Most studies on mental models of intelligent systems focused a single general mental model, e.g. [13,19,36]. Eiband et al. [13] highlighted the importance of identifying one "overarching user mental model" of a target group and indicated that within this model, several group-specific mental models may exist.

[1] For a detailed discussion on *folk theories*, e.g. see [15,16].

Indeed, some findings indicate a diverse landscape of mental models. In the domain of RS, Ghori et al. [19] showed that users mostly explain technical concepts, such as collaborative filtering, in their own words. In an interview study, Ngo et al. [36] revealed that mental models of RS might be technical or metaphorical. The study also suggests that users had different views on the importance of themselves in the recommendation process.

To summarize, while the elaboration of an overarching mental model for a system is useful, there is also strong support for the existence of diverse mental models within a population. To find a balance between one overarching mental model and an individual mental model for each user, we therefore argue to identify group-specific mental models. Even though qualitative approaches may provide some insights into the diversity of mental models, a quantitative approach is required to more precisely identify and classify these diverse models.

2.2 Methods for Eliciting Mental Models

Few studies have applied a quantitative approach to explore the mental models of intelligent systems. They mostly studied *effects* of mental models on the perception of a system. For instance, Kulesza et al. [26] induced different mental models and captured their "soundness" through multiple-choice questions. Thus, they did not directly investigate the structure and characteristics of mental models but the users' capacity of using them to simulate certain system outputs.

Other studies have used mixed-method approaches: Xie et al. [53] investigated the effects of mental model similarity on web page interaction performance in an experimental study. They combined a card sorting and a path diagram of web navigation and calculated different similarity measures based on these methods. A recent example studied mental models of cooperative AI agents in a game setting [18]. The researchers first applied a think-aloud task to explore the mental models. Then, a large-scale survey was conducted. We encourage such *informed quantitative studies* that exploit insights from former qualitative work.

Conceptual techniques, such as the repertory grid, pairwise rating, or card sorting [10,30] can be used to study mental models quantitatively. They are based on an existing body of concepts which needs to be explored before, e.g. through interviews. Thus, they do not rely on direct verbalization [10].

In repertory grid and pairwise rating, users rate different concepts on a certain scale or compare them with one another. This leads to a similarity matrix between the concepts representing the user knowledge. The data can be analyzed through e.g. multidimensional scaling [30]. While both methods have different advantages, they are either time-consuming or are limited in the number of concepts that can be studied. Therefore, Cooke [10] recommends to apply card-sorting techniques, if the number of concepts is higher than 25–30.

In card sorting, users assign certain cards, representing concepts, into categories. The method is often used in usability studies to determine navigation structures [9]. There are different types of settings: In closed card sorting, the content of the cards and the label as well as number of categories are fixed. In open card sorting, participants can label the cards themselves [9]. The method allows for the identification of common themes and differences in samples [44].

Table 1. Overview of the four general categories and their associated action cards we used in the card sorting task.

Acquisition of user data	*Comparing items or users*
[01] Recording my mouse clicks	[10] Comparing items regarding their content
[02] Asking me for my age	[11] Matching rating data of items
[03] Recording my dwell time on an item's detail page	[12] Calculating a similarity score between items
[04] Asking me to explicitly rate items	[13] Calculating a similarity score between users
Inference and aggregation	*Presenting recommendations*
[05] Determining my interest in item categories	[14] Suggesting items that are new to me
[06] Analyzing my current mood	[15] Showing items, I might like
[07] Combining all data about me to an abstract user profile	[16] Presenting items that other users liked in the past
[08] Adding additionally item data that users cannot see	
[09] Analyzing content of items	

3 Identifying Diversity and Commonalities of Mental Models

We developed a new setting based on card sorting. Card sorting is suitable for large online studies [4], allows open and closed settings [9], and can be used to include a wide range of concepts [10]. Our setting considers the subjectivity of mental models by providing a diverse range of pre-defined cards, and allowing participants to formulate their own thoughts using open cards and as many actions and steps as they find appropriate to describe their mental model.

In our card sorting setting, participants are presented with a set of cards representing typical RS actions and are asked to assign them to up to seven sequential steps. Our method assumes a procedural structure of the inner workings of a RS. This is in line with how these systems typically work and with observations in previous qualitative user studies [36,38,50]. The resulting card sorts of each participant represents their mental model of RS. Through hierarchical clustering, card sorts can be aggregated into groups, allowing us to characterize the differences and commonalities between mental models in a larger sample.

3.1 Cards Used as Actions of RS

We carefully created 35 cards for participants to express their mental model:

- 16 *action cards*, represent actions of the recommendation process (Table 1)
- 12 *distractor cards*, represent actions that are not part of the central recommendation process
- 4 *question mark cards*, provide the possibility to express uncertainty
- 3 *open cards*, let users express self formulated actions

The *action cards* correspond to typical paradigms used by RS while still "*speaking the language of the user*". We extracted concepts of mental models from former qualitative mental model studies [13,19,26,36,50] and contributed our own technical expertise on RS functioning. In particular, we followed the four general categories provided by Ngo et al. [36]: (1) *acquisition of user data*, (2) *inference and aggregation*, (3) *comparing items or users*, and (4) *presentation of recommendations*. For each category, we designed up to five cards (Table 1). We describe the rationale behind the action cards in the following.

(1) Acquisition of User Data (Cards 01–04): For any personalized RS, elicitation of user data and their preferences is a necessary prerequisite [42,46]. While these data can take various forms (e.g. ratings, purchases, clicks), the underlying concept appears to be well known by RS users and was mentioned in many in-depth qualitative user studies [13,19,36,50].

(2) Inference and Aggregation (Cards 05–09): In almost all cases RS do not perform their recommending on raw user data, but aggregate them or infer further (e.g. situational) data [1,3]. A similar concept can also be found in many user responses of prior interview studies. Users, for instance, mentioned (statistical) inference [50], or construction of a personal interest profile [19].

(3) Comparing Items or Users (Cards 10–13): Relating users or items is one of the most common techniques in RS design [25,37]. Such techniques, e.g. the commonly used *collaborative filtering*, are apparently well understood by users. In many prior mental models identified, the similarity between users or items was mentioned or played a central role [19,36].

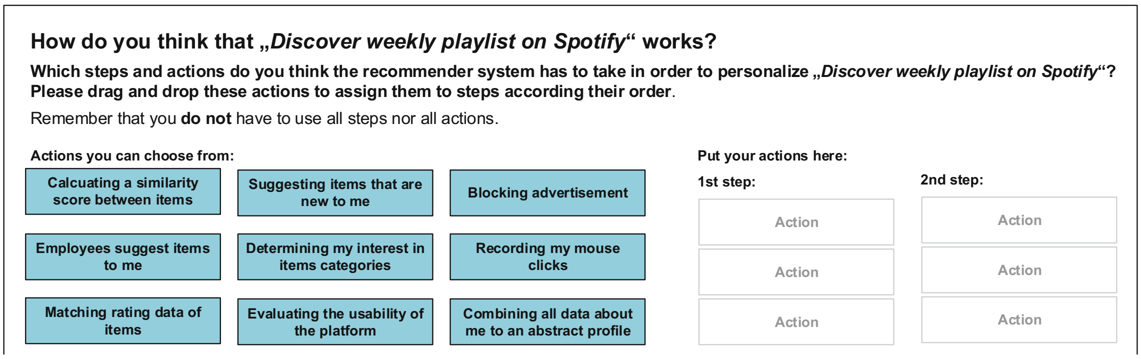
(4) Presenting Recommendations (Cards 14–16): While the form of presenting recommendations seems to play an inferior role in users' mental models [36], it is very relevant for RS research [23,48]. We thus decided to also include three actions for the presentation of RS outcome.

To further diversify answers and to enable analysis of the extent to which the mental models of participants diverge from a "*ground truth*", we added 12 *distractor cards*. These cards were chosen as misconceptions of RS as well as actions that are not part of the main personalization process. Distractor cards were collected by identifying such actions in results of a previous qualitative user study to which we had access (i.e. [36]). All distractor cards can be found in the supplementary material. Examples are: "*Employees suggest items for me*", "*Evaluating my satisfaction of recommendations*", and "*Blocking advertisement*".

Finally, we added *question mark* and *open cards*. The question mark cards account for uncertainties in participants' mental models, i.e. to indicate that there might be an unknown action performed in a certain step. Open cards account for any missing actions that were not part of the pre-labeled action or distractor cards, but are part of the subjective mental model.

4 User Study

Our online study consisted of three parts: instruction, mental model task, and measurement of technical knowledge and perception of RS. At the end, participants



How do you think that „*Discover weekly playlist on Spotify*“ works?

Which steps and actions do you think the recommender system has to take in order to personalize „*Discover weekly playlist on Spotify*“? Please drag and drop these actions to assign them to steps according their order.

Remember that you **do not** have to use all steps nor all actions.

Actions you can choose from:
Calcuating a similarity score between items
Suggesting items that are new to me
Blocking advertisement
Employees suggest items to me
Determining my interest in items categories
Recording my mouse clicks
Matching rating data of items
Evaluating the usability of the platform
Combining all data about me to an abstract profile
Put your actions here:
1st step:
Action
Action
Action
2nd step:
Action
Action
Action

Fig. 1. Excerpt of the mental model task with shortened description and examplified for *Discover weekly playlist on Spotify*. For reasons of space efficiency only nine actions and only two steps with three action slots are depicted here.

were debriefed and received 2.76 $ as compensation. We used Soscisurvey[2] as a survey platform in which we implemented the card sorting setting ourselves. On average participants took 13:39 minutes ($SD = 01{:}55$) to complete the study. This study was approved by the local ethics committee of the University of Duisburg-Essen. We included the complete lists of measures and items in the supplements. This section is organized according to the three parts of the user study.

4.1 Instruction

At the beginning, participants were presented with a definition of RS and the term “item”, which we defined as all content subject to recommendations, whether it is a product on Amazon, or a person suggested as friend on Facebook. Participants chose a RS they encounter regularly. Eight options were provided: *Top pics for you on Netflix*, *Video recommendations on YouTube*, *Discover weekly playlist on Spotify*, *Recommendations of similar items on Amazon*, *Friend recommendations on Facebook*, *Trending hashtags for you on Twitter*, *Personalized feed on Instagram*, and *Daily news recommendations on Google News*.

Additionally, participants could opt for “*None of the above*”, which resulted in an immediate end of this participant’s session. If any of the eight options was chosen, participants were instructed to keep the chosen RS and its items in mind as point of reference for all subsequent questions. As auxiliary reminder, their chosen RS was also explicitly displayed in several texts throughout the survey.

4.2 Mental Model Task

Next, participants had to complete the card sorting task described in Sect. 3. They were briefed to use their RS chosen in the previous part as reference while sorting their cards. All 35 cards were displayed on the left and participants were asked to sort as many of them as they deem appropriate via drag-and-drop in

[2] https://www.soscisurvey.de.

up to seven steps. The steps were displayed on the right (see Fig. 1). The open and question mark cards were shown at the bottom of the card list. Action and distractor cards were presented in a randomized order.

After the task, participants were asked about the *degree of fidelity* that reflected how well participants were able to express their mental model. We measured this using two self-created items on a 5-point Likert scale (1 ("I strongly disagree") to 5 ("I strongly agree")). The items were: "*I was able to express my ideas through the arrangements of steps and actions very well*" and "*I feel very certain about the arrangement of steps and actions.*", (Cronbach's $\alpha = .725$).

4.3 Measures

We asked participants about their perception of RS and technical knowledge: on the one hand, through self-created items on technical or metaphorical perception of RS, and, on the other hand, through standardized scales for social presence, trusting beliefs, transparency, and other user-centric measures of RS.

Perception of the RS: To assess whether participants perceived the chosen RS as rather technical or metaphorical, we included a self-created semantic differential consisting of twelve pairs such as "*machinelike*" vs. "*humanlike*" (Cronbach's $\alpha = .809$). Items were assessed on a 5-point Likert scale.

We used the social presence scale from Gefen and Straub [17] consisting of 5 items (e.g. "*There is a sense of human contact in the system.*"). Furthermore, we assessed *trusting beliefs* using items from McKnight et al. [32]. Trusting beliefs consist of three dimensions: benevolence, integrity, and competence. For all of these scales, items were rated on a 7-point Likert scale.

We measured *transparency*, *control*, and *perceived usefulness* using the *ResQue* inventory [41], and added *recommendation quality* and *perceived system effectiveness* from Knijnenburg et al. [24]. All items were assessed on a 5-point Likert scale.

Technical Knowledge of RS: We assessed the prior *technical knowledge* of participants by using three self-created items, e.g. "*In the past I learned about how recommender systems work*" (Cronbach's $\alpha = .818$). Additionally, we specifically asked for the *confidence* in the capability of learning about RS through one item, ("*I would be capable of understanding the recommendation process, if someone would explain it to me.*"). All items were measured on a 5-point Likert scale.

4.4 Participants

In total, 170 participants were recruited through the UK-based crowd-working platform *Prolific*[3]. Participants' age ranged from 18 to 67 ($M = 31.42$, $SD = 11.64$). Regarding gender, 71 participants identified as male and 99 as female.

[3] https://www.prolific.co/.

The sample was rather educated with 75 participants (44.1 %) holding a bachelor's degree, 55 participants (32.4 %) holding a high school diploma, and 26 (15.3 %) a master's degree. Six participants (3.5 %) held a PhD, while three participants (1.8 %) reported to hold less than a high school diploma. Five participants (2.9 %) indicated other degrees. Participants reported to have a low to moderate technical knowledge on RS ($M = 1.87$, $SD = .99$).

Generally, participants were able to express their mental model through the task well: Descriptive analysis revealed a moderate degree of fidelity with a mean score of 3.18 ($SD = 0.90$). Question mark cards were used very rarely (on average participants used $M = 0.03$ ($SD = 0.06$) of them). Only few open cards[4] were used: Participants created 63 cards themselves accounting for 2.32 % of all cards used. Most of them indicated similar ideas as existing action cards, e.g. "*Collecting other data such as gender*", or were specific to the RS, e.g. "*Monitoring what I watch*". 25.40 % of them were left blank or were unclear in their meaning.

Overall, participants used $M = 15.74$ ($SD = 8.20$) cards and $M = 4.90$ ($SD = 1.76$) steps to represent their mental model. When comparing action cards and all other cards, a t-test for paired samples revealed that action cards ($M = 9.85$, $SD = 4.15$) were used significantly more often than the others ($M = 5.88$, $SD = 4.80$), $t(169) = 14.94$, $p = .001$. The proportion of actions to distractors was at 26.61 % ($SD = 12.70$ %) on average, i.e. for each four action cards that were used in the mental model task, one was a distractor.

5 Results

We followed a *data-driven* approach to answer our RQs. Hierarchical clustering on participants' card sorts revealed three distinct user groups in our data. We conducted a descriptive analysis to compare the perceptions of RS of these groups. For the analyses we used SPSS 25 and R 4.0.2.

5.1 RQ1: Which Different Mental Models Do Users Hold Across RS?

To determine clusters among the different mental models expressed, we first calculated dissimilarities between card sorts. While card sorts are commonly evaluated this way, we faced two specific challenges in our task setting: (1) The order of steps, the cards were sorted in, was relevant to us, which is not taken into account by typical dissimilarity measures (e.g. *Jaccard Index*). (2) Participants were free to use any number of steps (up to a maximum of 7) and any number of cards (up to a maximum of 35), which resulted in many missing values. To overcome these two challenges, we calculated the dissimilarity between participants as follows:

$$dis(p,u) = 0.7 * d(p,u) + 0.3 * q(p,u)$$

[4] Note that a qualitative in-depth analysis of open cards was not within the scope of this work.

Table 2. Overview of user groups descriptive statistics. SI values refer to the cluster cut within each group.

Group	N	No. of cards	No. of steps	Degree of fidelity	No. of clusters	SI
		$M(SD)$	$M(SD)$	$M(SD)$		
1	66	8.53 (2.56)	4.09 (1.84)	3.34 (0.88)	4	0.298
2	79	16.76 (3.39)	4.99 (1.42)	3.46 (0.85)	2	0.238
3	25	31.60 (3.20)	6.76 (0.52)	2.82 (0.99)	7	0.132

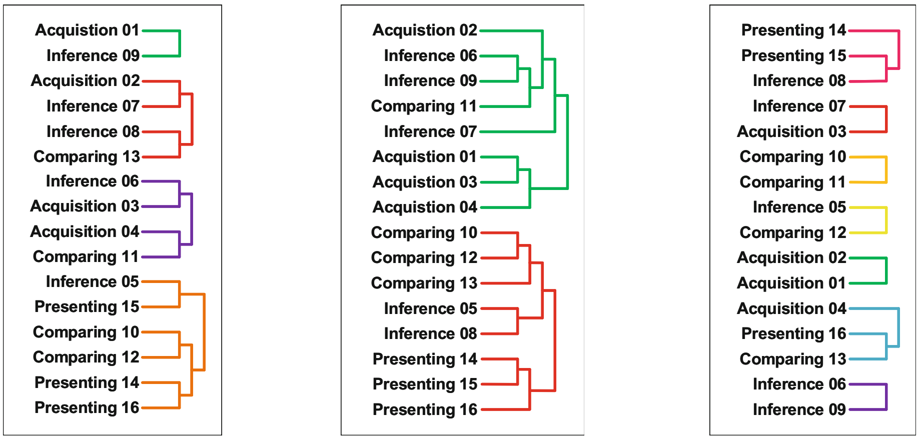
This dissimilarity calculation is based on two components. The first one ($d(p, u)$) determines the normalized *Manhattan distance* between any two participants. We interpret each participant's card sort as vector $p \in \mathbb{N}^c$, where c is equal to the number of available cards[5]. Each position of p, thus, corresponds to a specific card, while the value indicates the number of the step this participant assigned the card to. The Manhattan distance between these vectors accounts for challenge (1) as it considers the order of steps cards are sorted in. While this could be achieved with other similarity measures (e.g. *Euclidean distance*), the Manhattan distance treats coordinates as discrete, thus matching the discrete steps of our task design. This first component only includes cards that were used by both participants. Therefore, to account for (2), we add a second component ($q(p, u)$) as the difference of how many cards both participants used.

We acknowledged that both components should not equally contribute to the dissimilarity and deemed the step order as more important than the number of cards each participant used. Thus, we assigned different weights to each component and chose a factor of 0.7 for the first, and a factor of 0.3 for the second component. Detailed description of the formulas is included in the supplement.

Hierarchical Clustering. Hierarchical clustering can follow a *divisive* or an *agglomerative* clustering algorithm. Divisive clustering follows a top-down pattern, which starts with one cluster containing all items and divides them iteratively until each cluster contains only one single item. Agglomerative clustering takes the opposite approach and starts with each item as an own cluster and iteratively combines them until only one cluster remains [22].

We compared the *clustering coefficients* of divisive and agglomerative variants. This coefficient "describes the strength of the clustering structure" [22]. A coefficient closer to 1 indicates a stronger cluster structure and a better fit with the data. In our case, agglomerative clustering in tandem with the *Ward's* criterion [35,51] resulted in the best performance with a *clustering coefficient* of .949. To determine the number of clusters that fits the data best, we then compared cuts of the hierarchy at 2–7 clusters. For this, we used the respective *average silhouette index* (SI) [43] which reflects the *cohesion* within clusters

[5] We ignored open and question mark cards, since they cannot be compared easily.



(a) Group 1. (b) Group 2. (c) Group 3.

Fig. 2. Dendrograms depicting clusters of how cards have been sorted for each identified user group. Clusters and actions within are ordered regarding the median of steps, they have been sorted in.

and *separation* between clusters. The index ranges from −1 to 1. We found the highest SI of .237 when cutting at 3 clusters, and thus, 3 user groups.

Subsequently, we performed hierarchical clustering again. This second clustering was applied to cards within each of the 3 user groups and resulted in 2–7 clusters, depending on the group (Fig. 2 and Table 2). Below we describe the mental models of each user group in detail.

Group 1: Users with a Parsimonious Concept-Based Mental Model. This group used the lowest number of cards and steps (Table 2). Participants were convinced of their card sorts (degree of fidelity). Compared to the other groups, they expressed less prior knowledge in RS, but felt confident in understanding them. The group perceived RS as rather rational, planned, and machine-like (Table 3). The dendrogram of this group shows four major clusters (Fig. 2a) and that this group held a rather *concept-based* mental model.

The first major cluster is small and pertains to elicitation and analysis of implicit and less tangible user data ("recording of mouse clicks" (card 01) and "analyzing content of items" (card 09)) which can be considered as a starting point of RS processes. The second major cluster refers to the *inference of a user model* comprising of several processes including processes of data acquisition, inference, and comparison, all regarding the user (card 02, 07, 08, 13).

Following this, the third major cluster (card 05, 10, 12, 14–16) represents the *processing of the user model*. Further user data, e.g. "mood" and "dwell times", are analyzed and recorded. Additionally, the user data is connected to item data. In contrast to this, the fourth major cluster (card 03, 04, 06, 11) focuses clearly on the *processing of items*. It includes different processes, i.e. inference, comparison, and presentation of items.

In sum, this group was parsimonious in their use of cards, i.e. they only used few actions and steps. Participants of this group focused on the concepts of *the user model*, *the user model processing*, and on the *items*. In each cluster,

Table 3. Overview of perception of RS for each user group.

Variables		$M(SD)$	
	Group 1	Group 2	Group 3
Technical knowledge of RS			
Knowledge of RS	1.56 (0.83)	1.97 (1.03)	2.20 (1.10)
Confidence	4.23 (0.74)	4.30 (0.65)	3.72 (0.84)
Perception of RS			
Technical/ metaphorical	2.74 (0.48)	2.65 (0.62)	3.01 (0.74)
Social presence	3.11 (1.44)	3.29 (1.58)	3.50 (1.62)
Transparency	3.76 (0.79)	4.09 (0.75)	3.88 (0.78)
Trusting beliefs (TB)	3.85 (1.22)	4.04 (1.29)	4.20 (1.38)
TB benevolence	3.39 (1.42)	3.44 (1.60)	3.95 (1.47)
TB integrity	3.61 (1.36)	3.66 (1.50)	4.18 (1.54)
TB competence	4.45 (1.48)	4.88 (1.34)	4.41 (1.57)

processes are mixed (e.g. acquisition and inferences processes regarding the user model, comparisons, inference, and presentation regarding the items).

Group 2: Users with a Feasible Procedural Mental Model. This group could express their mental model through the task well. Their prior knowledge was higher than in group 1, but lower than in group 3 (see Table 3). Like group 1, they perceived RS as rational, planned, machine-like, but as more transparent. The card sorting task resulted in two major clusters (Fig. 2b).

The first major cluster can be divided into two sub-clusters. The first one (card 02, 06, 07, 09, 11) pertains to the inference of a user model using *contextual* user data, such as the "age" or "mood" of the user. The second sub-cluster (card 01, 03, 04) pertains to the acquisition of *interaction* data that is dependent on the use of the RS, e.g. "mouse clicks", "dwell time".

The second major cluster consists of three sub-clusters that represent different processes of RS: comparison of items and users (cards 10, 12, and 13), inferences of the user's interest based on items (card 05, 08), and finally the presentation of recommendations (card 14–16).

In sum, this group showed a procedural mental model that reflected our proposed procedure best (Sect. 3 and Table 1). Only the first major cluster represented a more nuanced understanding of the acquisition of user data which differed from our proposed procedure. Group 2 views the user model as a starting point that is characterized by contextual data, i.e. data that exist prior to the interaction with RS. Thus, they distinguish between contextual and interaction data. The second major cluster represented the last three steps of our proposed procedure accordingly.

This user group seemed to have the most structured comprehension of RS which is indicated by the rather high values for the degree of fidelity, confidence, and transparency.

Group 3: Users with an Extensive Social-Focused Mental Model. This group used the highest number of cards and steps (Table 2). The confidence in their card sorts was the lowest, but they expressed a higher knowledge about RS[6]. Group 3 perceived the RS as more empathetic, spontaneous, and human-like. We found this tendency as well when examining the values for social presence and trusting beliefs, which were the highest in this group (Table 3). The dendrogram (Fig. 2c) reveals seven major clusters of action cards.

The first major cluster mainly contains the presentation of recommendations, showing items that are new (card 14) and the user might like (card 05). Based on these presented items, additional, invisible data are considered (card 08). The second major cluster combines user information to an abstract profile (card 07), to which user data (e.g. "dwell time", card 03) are added. The third and fourth major cluster mostly pertain to comparison processes of the items (card 10–12) and determination of user interests (card 05).

The fifth cluster refers to the data acquisition of explicit user data (card 01, 02), while the sixth pertains to items in relation to users (e.g. user ratings of items (card 04), similarity between users (card 13), and presentation of what users liked in the past (card 16)). These processes were mostly assigned to step 3. The last cluster refers to inference processes on the "mood that the user is currently in" (card 06) and "analyzing content of items" (card 09).

This group used nearly all cards and steps, i.e. many distractor, open, and question mark cards were used. We conclude that this user group might have an extensive mental model consisting of many different processes that go beyond the recommendation process described in Sect. 3.

In sum, the mental model of group 3 appears rather unstructured. This is reflected by the high number of clusters (i.e. many small unrelated islands). Like group 1, participants of this group seem to follow a rather concept-based mental model. Yet, they distinctively assigned more human attributes and social presence to the system indicating a higher social focus of their mental models.

5.2 RQ2: How Do These Mental Models Relate to the Perception of RS?

Due to the *exploratory* nature of our approach, we analyzed our results descriptively[7]. First, we explored if we find differences for the RS choice in the measures. The descriptive data revealed that confidence intervals (CI) of means of each chosen RS largely overlap for all measures. This indicates that the results are independent of the particular RS a participant had in mind. Instead, we conclude that measured differences resulted from the particular mental model a participant held.

Then, we analyzed the group differences based on 95 % CI of mean differences and effect sizes (using Cohen's *d* with pooled standard deviation to account for different group sizes). To this end, we first performed a visual analysis of CI of

[6] However, the level of knowledge in all groups can be considered as low to moderate.
[7] An overview of all descriptive data can be found in the supplement.

Table 4. Overview of descriptive analysis.

Group comparisons	Cohen's *d*	95% CI	Mean diff.	95% CI
Confidence				
Group 1 vs. 3	−.66	[−1.13, −0.19]	.51	[0.10, 0.91]
Group 2 vs. 3	−.83	[−1.29, −0.37]	.58	[0.19, 0.98]
Technical/ metaphorical				
Group 1 vs. 3	.48	[0.01, 0.95]	−.27	[−0.60, 0.07]
Group 2 vs. 3	.55	[0.10, 1.01]	−.36	[−0.69, −0.03]
Transparency				
Group 1 vs. 2	.43	[0.10, 0.76]	−.33	[−0.64, −0.02]
Knowledge of RS				
Group 1 vs. 2	.43	[0.10, 0.77]	−.41	[−0.81, −0.02]
Group 1 vs. 3	.70	[0.23, 1.17]	−.64	[−1.19, −0.09]

group means for each measure. We only report results with moderate to large effect sizes and CI with little or no overlap (Table 4).

Regarding *confidence*, we found that group 3 was less confident than group 1 and 2. This indicates that users with a social-focused mental model were less confident in their capabilities to understand the RS. The analysis revealed the same pattern regarding *technical vs. metaphorical perception* suggesting that group 3 tended to view RS as more human-like than the other two groups. Concerning *transparency*, we found a difference between group 1 and 2 indicating that the procedural mental model might be associated with higher transparency perception. Finally, regarding *knowledge of RS*, we found that group 1 expressed lower knowledge than group 2 and 3.

The descriptive analysis suggests that the precision of the measures were low. Therefore, the results give first indications of relevant relationships between the structure of mental models and RS perceptions.

6 Discussion

This work extends the existing research body on the measurement of mental models through a novel card sorting setting. While it does not investigate single mental models in detail, as fully qualitative methods would, our approach allows for relevant analytical insights. We analyzed *the diversity of mental models* in a large sample. Thus, we envision our card sorting setting as beneficial in a second research stage, after a first general mental model was already revealed.

In line with prior work of Norman [38], who observed the transfer of mental models from one system to another, we found that mental models exist *across systems*. Interestingly, we did not find any relationship between the referenced RS and the perception of RS, i.e. they were independent of another. In fact, differences in users' perceptions of RS were only dependent on users' mental model.

We conclude that mental models appear to be more critical for the perception of RS than the system itself. Hence, for contemporary user-centered design of RS, we suggest a shift from system-focused to mental-model-focused research.

In the following, we discuss the mental models of RS and their relation to the perception of RS. Furthermore, we address RQ3 (*Based on the identified mental models, which implications can be derived from them for the design of transparent intelligent systems?*) and discuss practical implications for the development of more user-friendly, trustworthy, and transparent user interfaces.

6.1 Seeing Is Not Understanding

Many participants perceived the referenced RS (e.g. Youtube, Netflix, Spotify) as transparent. However, the transparency perception cannot be ascribed to a factual knowledge about the inner workings of these RS. Firstly, because these systems do not provide any sophisticated explanatory components and, secondly, because participants reported a low to moderate technical expertise of RS. We therefore attribute the transparency perceptions to participants' mental models, which are based on subjective explanations of how the RS work. These explanations, hence, merely form an *impression* of understanding that may not match the actual systems' functioning. In other words, "*seeing*" a system does not necessarily translate to *understanding* it [2]. We argue that such mental models, based on vague information of how the system works, may result in a gap between actual system behavior and users' expectations—a concept known as *gulf of evaluation* [39]. Such gulf was observed to result in false assumptions and erroneous behavior [33,34]. Morris [33] found that social media users can misinterpret the opaque algorithms responsible for composing their news feed. In the case of Morris' observation, this led to the negative public misperception that new mothers post excessively about their newborns when in fact they do not. Muramatsu and Pratt [34] could show that false assumptions and erroneous mental models can be corrected through transparency.

Practical Implication: Dare to Provide Transparency to Users. To avoid a false sense of understanding a system, typical straightforward explanatory components might be too shallow to provide "real" transparency in terms of an actual user comprehension. Thus, users' mental models need to be regarded, evaluated, and, if flawed, corrected by providing factually accurate insights into the system's inner working. While we note that such a correction could benefit from knowing the active user's mental model during runtime, it could also be based on a general elicitation of mental models prevalent in a user base. The presented study demonstrates how such elicitation could be performed. Yet, we acknowledge that further research in eliciting mental models and providing transparency of RS is necessary as intelligent systems become increasingly sophisticated.

Previous research has indicated users' interest in more algorithmic transparency, e.g. [15,31]. Our study extends on that: It highlights that there is not only a user *interest*, but also that users feel *confidence* in their ability to understand intelligent systems when appropriate explanations are present. This is especially interesting considering the low technical knowledge of our participants. We, thus, encourage developers to dare to provide sophisticated components of transparency, e.g. in form of explanations [7,21] or visualizations [5,28,29].

6.2 Procedural vs. Concept-Based Mental Models

We could uncover three different mental models of RS that coexist in a large sample of RS users. We observed that these models exhibited different structures and perceptions of RS. Concept-based and procedural mental models were the most prevalent models that co-existed in our sample. An extensive and social-focused mental model was held by a minority of the participants.

The mental model of group 2 reflected the procedure, that our method was based on, best. Due to the opacity of RS, we cannot claim this procedure to be a *ground truth* of RS. Yet, it is based on established publications of researchers and practitioners in the field of RS and we deem it—to a certain degree—accurate. In this regard, group 2, interestingly, felt the highest degree of fidelity in expressing their mental model through the card sorting. Based on this, we assume that the mental model of this group was rather well-defined. Therefore, they perceived the highest transparency of RS. The well-defined mental model might also be the cause for the highest competence perception: The RS was perceived as reasonable leading to comprehension of the system and appreciation of its competence. As this group expressed low technical knowledge of RS, we conclude a close connection between a well-defined mental model, understanding the actual system functioning, the transparency and competence perception of the system.

Group 1 and 3 did not strongly adhere to a process-based mental model. It seems that they did not use the steps in a chronological, but in a concept-wise manner. Inspection of the clusters in the dendrogram of group 1 (Fig. 2a) showed that many clusters consisted of actions from different chronological stages. The second cluster, for instance, comprised of four cards (02, 07, 08, 13) of which three cards belong to another chronological stage. Yet, they shared a conceptual focus: the user model. The most frequent and strict concepts in the mental models of group 1 and 3 were *item- vs. user-based recommending.*

Practical Implication: Increase Transparency Through Procedural Explanations. We conclude that there are several perspectives on a RS that users can adopt. Delivering different user interfaces to each of these groups might address this issue best. For users adhering to a procedural mental model, explanations that emphasize the chronology of the recommendation process can be useful. To prevent aforementioned false assumptions, we suggest great care that explanations reflect the actual recommendation process as closely as possible.

Users that adhered to a concept-based mental model perceived lower transparency. Hence, we suggest explaining the concepts more clearly to those users, i.e. practitioners could provide clear definitions and examples of explicit and implicit user data and explain their application in RS. Similarly, practitioners can stress clearly whether users or items are compared to generate recommendations. The latter was recently identified to cause confusion for users [36].

However, we acknowledge that treating each user group differently is not always possible, e.g. when no information on the active user is available. While our quantitative approach could be used to correlate mental models to user interaction data (e.g. mouse movements), thus forming a baseline for inferring the user's mental model during runtime, this demands further studies. Yet, in our study, we identified some procedural aspects in all user groups and are thus confident that a procedural perspective could be "imposed" on users with a more concept-wise mental model. Hence, we recommend considering procedural explanations in RS. Apart from matching most user expectations, our findings suggest that this form of explanation also results in a higher perceived transparency.

6.3 Technical vs. Humanized RS

While group 1 and 2 held a rather technical understanding of RS (rational, and machinelike), group 3 described them as neutral to metaphorical (empathetic, spontaneous, and humanlike). Thus, group 3 *humanized* the RS more than the other groups, i.e. they ascribed humanlike characteristics to a non-human agent. This humanization acts as a mechanism to combat uncertainty and situations in which a system seems unpredictable [14]. This effect might be at work here: Besides the more humanized mental model compared to the other groups, group 3 expressed low confidence in the ability to learn about the system.

Prior work in autonomous vehicles has indicated a link between humanization and more trust in the non-human agent [52]. Our study shows that this mechanism might also occur in intelligent systems: group 3 perceived higher levels of trusting beliefs. Furthermore, descriptive values indicate a higher social presence for group 3. We ascribe this also to the more metaphorical and humanized mental model of this group. In line with prior work [8,27], this social presence may act as mediator between humanization and trusting beliefs in group 3.

Practical Implication: Educate Users and Create Social Presence. Uncertain users might hold an unstructured mental model including metaphorical concepts. As a consequence, such user groups might perceive the system as unpredictable and tend to humanize it. From this, we derive two implications for practitioners: (1) There is a need to educate uncertain users, so that they do not need to develop metaphorical or humanized mental models. As a result the system could be perceived as more predictable and transparent. Yet, we also note that some desirable aspects may arise from a higher social presence of RS and thus, (2), suggest to include social aspects into a user interface. This could, for instance, be realized by adding elements that express metaphors or using a metaphorical language. We, however, note that this is speculative and emphasize the necessity of investigating these aspects in greater depth.

6.4 Limitations

We created the cards as carefully as possible and added open cards to formulate new actions. Still, some actions that participants created were redundant with our pre-formulated cards. Therefore, we assume that some participants did not read all cards or did not fully understand them. Thus, we deem 35 cards as maximum in such settings and reconsider wording choices. Another limitation of our study concerns our task setting. While participants were able to express procedural mental models well, this did not necessarily apply to other forms of mental models (although participants managed to express them anyway, see Sect. 6.2). We conclude that the task design could be slightly adjusted to, for instance, express parallel actions or feedback loops. This could, for instance, be achieved through concept networks or flow diagrams. We also acknowledge that we have included only a small fraction of all existing RS in our study and that RS represent only one facet of the full range of intelligent systems. Future work might investigate mental models of additional RS and other intelligent systems.

7 Conclusions and Future Work

We introduced a method that enables us to identify mental models quantitatively and to examine their diversity in large samples and across platforms. It poses a substantial extension of prior research on mental models of intelligent systems which relied on qualitative studies with small samples.

We could reveal a relation between mental model structures and user perception of RS: Procedural mental models were positively related to transparency, implying that transparency can be increased through procedural explanations. Such type of explanations could also be imposed on users who hold a concept-based mental model. Additionally, uncertain users might hold social-focused mental models and perceive RS as more humanlike, which leads to ambivalent results: While social-focused mental models might positively relate to trust, they might lead users to be less confident and perceive a system as unpredictable.

Finally, this study highlights that mental models exist *across systems*, i.e. the perception of RS mainly depends on the mental models, and not on the particular system. We consequently emphasize the relevance of mental models for designing user-friendly intelligent systems and advocate a shift from system-focused to mental-model-focused research in that area.

Our method allows to identify mental model in statistically representative user studies, and thus, to make generalizable inferences about the mental models in a target audience and their relations to system perceptions. Moreover, we suggest an analysis of the relationship between user characteristics (e.g. personality traits such as need for cognition) and mental models of intelligent systems. Our method could be used to identify user groups that relate to certain personality profiles. This could contribute to measuring a user's mental model during runtime, enabling presentation of personalized transparency components, tailored towards their mental model and personality. This might be especially useful for system applications that require long-term relation between a user and a system.

References

1. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 217–253. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-85820-3_7
2. Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
3. Beliakov, G., Calvo, T., James, S.: Aggregation Functions for Recommender Systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 777–808. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_23
4. Bussolon, S., Russi, B., Missier, F.D.: Online card sorting: as good as the paper version. In: Proceedings of the 13th Eurpoean Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical Systems. ECCE 2006, New York, NY, USA. ACM (2006). https://doi.org/10.1145/1274892.1274912
5. Cardoso, B., Brusilovsky, P., Verbert, K.: Intersectionexplorer: the flexibility of multiple perspectives. In: Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems. IntRS 2017, CEUR Workshop Proceedings, pp. 16–19 (2017)
6. Castelo, N., Bos, M.W., Lehmann, D.R.: Task-dependent algorithm aversion. J. Mark. Res. **56**(5), 809–825 (2019). https://doi.org/10.1177/0022243719851788
7. Cheng, H.F., et al.: Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In: Proceedings of the 2019 Conference on Human Factors in Computing Systems. CHI 2019, New York, NY, USA, pp. 559:1–559:12. ACM (2019). https://doi.org/10.1145/3290605.3300789
8. Choi, J., Lee, H.J., Kim, Y.C.: The influence of social presence on evaluating personalized recommender systems. In: Pacific Asia Conference on Information Systems, p. 49. AISeL (2009)
9. Conrad, L.Y., Tucker, V.M.: Making it tangible: hybrid card sorting within qualitative interviews. J. Doc. **75**(2), 397–416 (2019). https://doi.org/10.1108/JD-06-2018-0091
10. Cooke, N.J.: Varieties of knowledge elicitation techniques. Int. J. Hum.-Comput. Stud. **41**(6), 801–849 (1994). https://doi.org/10.1006/ijhc.1994.1083
11. Cramer, H., et al.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Model. User-Adapted Inter. **18**(5), 455 (2008). https://doi.org/10.1007/s11257-008-9051-3
12. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. **144**(1) (2015). https://doi.org/10.1037/xge0000033
13. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: 23rd International Conference on Intelligent User Interfaces. IUI 2018, pp. 211–223. ACM (2018). https://doi.org/10.1145/3172944.3172961
14. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. **114**(4), 864–886 (2007). https://doi.org/10.1037/0033-295X.114.4.864

15. Eslami, M., Vaccaro, K., Lee, M.K., Elazari Bar On, A., Gilbert, E., Karahalios, K.: User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In: Proc. of the 2019 Conference on Human Factors in Computing Systems. CHI 2019, New York, NY, USA, p. 1–14. ACM (2019). https://doi.org/10.1145/3290605.3300724
16. French, M., Hancock, J.: What's the folk theory? reasoning about cyber-social systems (2017). https://ssrn.com/abstract=2910571, https://doi.org/10.2139/ssrn.2910571
17. Gefen, D., Straub, D.W.: Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. Omega **32**(6), 407–424 (2004). https://doi.org/10.1016/j.omega.2004.01.006
18. Gero, K.I., et al.: Mental models of AI agents in a cooperative game setting. In: Proceedings of the 2020 Conference on Human Factors in Computing Systems, Honolulu HI USA, pp. 1–12. ACM, April 2020. https://doi.org/10.1145/3313831.3376316
19. Ghori, M.F., Dehpanah, A., Gemmell, J., Qahri-Saremi, H., Mobasher, B.: Does the user have a theory of the recommender? A pilot study. In: Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2019), Copenhagen, DK, p. 9. ACM, September 2019
20. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. CSCW 2000, New York, NY, USA, pp. 241–250. ACM (2000). https://doi.org/10.1145/358916.358995
21. Hernandez-Bocanegra, D.C., Donkers, T., Ziegler, J.: Effects of argumentative explanation types on the perception of review-based recommendations. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 20 Adjunct, New York, NY, USA, pp. 219–225. ACM (2020). https://doi.org/10.1145/3386392.3399302
22. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. Wiley, Hoboken (1990). https://cds.cern.ch/record/1254107
23. Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 309–352. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_9
24. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model. User-Adap. Inter. **22**(4), 441–504 (2012). https://doi.org/10.1007/s11257-011-9118-4
25. Koren, Y., Bell, R.: Advances in Collaborative Filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 77–118. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_3
26. Kulesza, T., Stumpf, S., Burnett, M., Kwan, I.: Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In: Proceedings of the 2012 Conference on Human Factors in Computing Systems. CHI 2012, Austin, Texas, USA, pp. 1–10. ACM (2012). https://doi.org/10.1145/2207676.2207678
27. Kunkel, J., Donkers, T., Michael, L., Barbu, C.M., Ziegler, J.: Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the 2019 Conference on Human Factors in Computing Systems. CHI 2019, New York, NY, USA, pp. 1–12. ACM (2019). https://doi.org/10.1145/3290605.3300717

28. Kunkel, J., Loepp, B., Ziegler, J.: 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 2017, New York, NY, USA, pp. 3–15. ACM (2017). https://doi.org/10.1145/3025171.3025189
29. Kunkel, J., Schwenger, C., Ziegler, J.: Newsviz: depicting and controlling preference profiles using interactive treemaps in news recommender systems. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP 2020, New York, NY, USA, pp. 126–135. Association for Computing Machinery (2020). https://doi.org/10.1145/3340631.3394869
30. Langan-Fox, J., Code, S., Langfield-Smith, K.: Team mental models: techniques, methods, and analytic approaches. Hum. Factors **42**(2), 242–271 (2000). https://doi.org/10.1518/001872000779656534
31. Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proc. of the 11th International Conference on Ubiquitous Computing. UbiComp 2009, New York, NY, USA, pp. 195–204. ACM (2009). https://doi.org/10.1145/1620545.1620576
32. McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and validating trust measures for e-commerce: an integrative typology. Inf. Syst. Res. **13**(3), 334–359 (2002). https://doi.org/10.1287/isre.13.3.334.81
33. Morris, M.R.: Social networking site use by mothers of young children. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW 2014, New York, NY, USA, pp. 1272–1282. ACM (2014). https://doi.org/10.1145/2531602.2531603
34. Muramatsu, J., Pratt, W.: Transparent queries: investigation users' mental models of search engines. In: Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval. SIGIR 2001, New York, NY, USA, pp. 217–224. ACM (2001). https://doi.org/10.1145/383952.383991
35. Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J. Classif. **31**(3), 274–295 (2014). https://doi.org/10.1007/s00357-014-9161-z
36. Ngo, T., Kunkel, J., Ziegler, J.: Exploring mental models for transparent and controllable recommender systems: a qualitative study. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Genoa Italy, pp. 183–191. ACM, July 2020. https://doi.org/10.1145/3340631.3394841
37. Ning, X., Desrosiers, C., Karypis, G.: A comprehensive survey of neighborhood-based recommendation methods. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 37–76. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_2
38. Norman, D.A.: Some Observations on Mental Models. In: Gentner, D., Stevens, A.L. (eds.) Mental Models, pp. 7–14. Psychology Press, New York (1983)
39. Norman, D.A.: The Design of Everyday Things. Basic Books Inc., New York (1988). ISBN 978-0-465-06710-7
40. Prahl, A., van Swol, L.: Understanding algorithm aversion: When is advice from automation discounted? J. Forecast. **36**(6), 691–702 (2017). https://doi.org/10.1002/for.2464
41. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM Conference on Recommender Systems - RecSys 2011, Chicago, Illinois, USA, p. 157. ACM (2011). https://doi.org/10.1145/2043932.2043962

42. Ricci, F., Rokach, L., Shapira, B.: Recommender systems: introduction and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 1–34. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_1
43. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
44. Rugg, G., McGeorge, P.: The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. Expert. Syst. **14**(2), 80–93 (1997). https://doi.org/10.1111/1468-0394.00045
45. Rumelhart, D.E., Norman, D.A.: Representation in Memory. No. 116 in CHIP report, University of California, San Diego (1983)
46. Sparling, E.I., Sen, S.: Rating: how difficult is it? In: Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys 2011, New York, NY, USA, pp. 149–156. ACM (2011). https://doi.org/10.1145/2043932.2043961
47. Torkamaan, H., Barbu, C.M., Ziegler, J.: How can they know that? A study of factors affecting the creepiness of recommendations. In: Proceedings of the 13th ACM Conference on Recommender Systems. RecSys 2019, New York, NY, USA, pp. 423–427. ACM (2019). https://doi.org/10.1145/3298689.3346982
48. Tsai, C.H., Brusilovsky, P.: Beyond the ranked list: User-driven exploration and diversification of social recommendation. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces. IUI 2018, New York, NY, USA, pp. 239–250. ACM (2018). https://doi.org/10.1145/3172944.3172959
49. Tsai, C.H., Brusilovsky, P.: Explaining recommendations in an interactive hybrid social recommender. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI 2019, New York, NY, USA, pp. 391–396. ACM (2019). https://doi.org/10.1145/3301275.3302318
50. Tullio, J., Dey, A.K., Chalecki, J., Fogarty, J.: How it works: a field study of non-technical users interacting with an intelligent system. In: Proceedings of the 2007 Conference on Human Factors in Computing Systems. CHI 2007, New York, NY, USA, pp. 31–40. ACM (2007). https://doi.org/10.1145/1240624.1240630
51. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963). https://doi.org/10.1080/01621459.1963.10500845
52. Waytz, A., Heafner, J., Epley, N.: The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. **52**, 113–117 (2014). https://doi.org/10.1016/j.jesp.2014.01.005
53. Xie, B., Zhou, J., Wang, H.: How Influential are mental models on interaction performance? Exploring the gap between users' and designers' mental models through a new quantitative method. Adv. Hum.-Comput. Inter. **2017**, 1–14 (2017). https://doi.org/10.1155/2017/368354
54. Yang, R., Shin, E., Newman, M.W., Ackerman, M.S.: When fitness trackers don't 'fit': End-user difficulties in the assessment of personal tracking device accuracy. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp 2015, New York, NY, USA, pp. 623–634. ACM (2015). https://doi.org/10.1145/2750858.2804269
55. Zhou, J., Chen, F.: 2D transparency space—bring domain users and machine learning experts together. In: Zhou, J., Chen, F. (eds.) Human and Machine Learning. HIS, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90403-0_1

The following article is reused from:

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12. ACM, 2019. doi: 10.1145/3290605.3300717

Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, Jürgen Ziegler
University of Duisburg-Essen, Duisburg, Germany
{firstname.lastname}@uni-due.de

ABSTRACT

Trust in a Recommender System (RS) is crucial for its overall success. However, it remains underexplored whether users trust personal recommendation sources (i.e. other humans) more than impersonal sources (i.e. conventional RS), and, if they do, whether the perceived quality of explanation provided account for the difference. We conducted an empirical study in which we compared these two sources of recommendations and explanations. Human advisors were asked to explain movies they recommended in short texts while the RS created explanations based on item similarity. Our experiment comprised two rounds of recommending. Over both rounds the quality of explanations provided by users was assessed higher than the quality of the system's explanations. Moreover, explanation quality significantly influenced perceived recommendation quality as well as trust in the recommendation source. Consequently, we suggest that RS should provide richer explanations in order to increase their perceived recommendation quality and trustworthiness.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Recommender Systems, Trust, Explanations, User Study, Structural Equation Modelling, Counterfactual Analysis

ACM Reference Format:
Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300717

1 INTRODUCTION

Contemporary online platforms typically rely on *impersonal* recommendation sources, i.e. automated Recommender Systems (RS), that automatically generate recommendations in order to faciliate users' decision making when facing a large number of alternatives. Even though recommendation algorithms have become highly accurate in terms of estimating a user's preferences [1, 15], they oftentimes appear as "black boxes" by concealing important details from their users. As a consequence, users create an unfitting mental model of the RS which may result in distrust and ultimately even in rejection of the system's recommendations [17, 46]. Hence, several researchers argue that especially the trustworthiness of a RS should be considered when assessing its quality [2, 22, 36].

RS are faceless entities lacking the human properties that are important for the development of trust, thus making it difficult for users to form bonds of any kind. One way to alleviate this, is to introduce social components into RS [6, 26]. There is a growing number of websites where automated and human-generated recommendations are combined—the latter, for example, in form of customer product reviews. For the reasons above, *personal* recommendation sources, i.e. users providing recommendations, are often associated with a higher trustworthiness [26, 44].

In the same line, designers of RS often strive to increase transparency and trustworthiness by providing textual explanatory components for recommendations [5, 46, 50]. A very common technique is to indicate similarity between recommendations and items the user is currently browsing or has expressed preferences for in the past. A well-known example for the former is Amazon's "Users who bought . . . also bought . . . " explanation. Similar kinds of explanations are applied by, for instance, Netflix and Spotify. Even though the effectiveness of such simplistic approaches utilizing similarity-based explanations has been questioned [5, 12], a thorough empirical comparison with systems using richer explanations—especially in terms of the perceived trustworthiness and its influential factors—is still missing.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05. . . $15.00
https://doi.org/10.1145/3290605.3300717

We argue that overly simplistic explanations lack the expressiveness and social properties that are relevant to establish trust in a recommendation source. In order to find empirical support for this assumption, we conducted a user study in which we let participants assess recommendations that were either selected by another person or by a typical RS. Additionally, the recommended items were accompanied by individually composed explanations in the personal condition or similarity-based ones for the RS. By utilizing tools of causal statistical inference, i.e. *structural equation modeling* [33] and the *counterfactual framework* [18, 35, 40], we were able to reveal that the richness of explanations plays a pivotal role in trust-building processes. Although, as compared to the RS, humans were usually less accurate in estimating preferences, the explanations for their choice were more elaborate and comprehensible such that the overall quality of recommendations was deemed to be equal.

As a consequence, it appears reasonable to develop RS towards incorporating explanatory components that imitate more closely the way humans exchange information. Counterfactual analysis helped us answer questions about a hypothetical situation in which RS would do so: As it turns out, without any changes to the underlying algorithm, replacing similarity-based explanations with human-like ones, the quality of recommendations can be expected to improve by around 13 %.

Moreover, up to now research on trust in RS has been concentrating predominantly on an initial perception of trust and little research addresses temporal development of trust in recommendation sources [e.g. 8, 49]. Participants in our study received two recommendations over the course of two weeks, which allowed us to assess trust development over time. While trust in humans remained constant, we could observe a slight decrease for their automated counterparts. Although not statistically significant, we assume systematic effects that tackle information asymmetry and, through this, unfulfilled expectations.

The contributions of this paper can be summarized as follows:

- We conducted a user study that compared personal to impersonal recommendation sources. It is shown that there exist differences between the groups in how recommendations are perceived and in how bonds are created towards the recommendation source.
- We structurally model direct and indirect effects between constructs of major interest for RS research. Concretely, we reveal complex dependencies between *explanation quality*, *recommendation quality*, *social presence*, and *trustworthiness.*
- We provide empirical evidence that simplistic explanations fall short in terms of their benefit for recommendations when compared to human explanations. We suggest that RS should be equipped with more sophisticated means of explaining their decisions. Natural language information exchange employed by humans should be the reference point.
- The contribution is also of theoretical value as we utilize profound statistical tools that allow for causal interpretation of effects. We argue that RS research will benefit substantially from this direction because it opens up an perspective that cannot be achieved with correlative studies.

The remainder of this paper is organized as follows. Section 2 examines relevant literature and puts them in relation. We describe our empirical study and the tools we used in Section 3 and present results in Section 4. Finally, implications of our findings are discussed in Section 5 and summarized in Section 6. The latter also addresses limitations and future work.

2 RELATED WORK

RS have become ubiquitous means that proactively filter information in order to help users find interesting items [38]. Providing recommendations not only helps users make decisions, thus reducing their cognitive load [19, 37], but also increase purchases and general user satisfaction [38]. Nearly all contemporary online platforms, such as Amazon, Netflix, and Facebook, make use of RS [14, 16, 43]. While for a long time research in RS focused primarily on algorithmic accuracy, it recently began to shift onto more user-centered qualities [4, 23, 27, 32] such as the degree of control [20], the transparency [46] and the trustworthiness [47] of a RS.

Trust in Recommender Systems

Trust is an important factor in human-machine interaction [28] and arguably of special interest for RS, since taking an advice is a highly trust-dependent behavior [30, 31]. Not surprisingly, increasing the trustworthiness of a RS has been shown to increase purchase volume [34, 46] and customer loyalty [46], among others.

From a cognitive science perspective, it is a non-trivial task to define what constitutes trust. Consequently, there are various definitions of trust in the literature. In this paper we follow McKnight et al. [30, 31] and their interdisciplinary model of trust. The model comprises four general constructs that are directly or indirectly influencing trust-related behavior: a user's *disposition to trust* together with their *institution-based trust*, *trusting beliefs*, and *trusting intentions*. The *disposition to trust* describes the trusting stance and trustfulness of a person, such as their general faith in humanity. In contrast to the rather constant *disposition to trust*, *institution-based trust* is ephemeral and lasts only for certain situations (e.g. visiting an online shop). *Disposition to trust* and *institution-based trust* together build the foundation for *trusting beliefs*. *Trusting* *beliefs* directly concern characteristics of the trustee, which are threefold in the model of McKnight et al.: *integrity* (the trustee's reliability and honesty), *benevolence* (the trustee's motives such as altruism and goodwill) and *competence* (the trustee's ability to fulfill the truster's needs). Before a person finally commits to a trust-related behavior (e.g. making an online purchase), *trusting intentions* need to be present. *Trusting intentions* itself consist of four subconstructs: *willingness to depend* (the general readiness to make oneself vulnerable to the trustee), *follow advise* (the intention to take an advice of the trustee), *give information* (the willingness to share some private information with the trustee) and *make purchase* (the intention to actually purchase something). Trusting intentions highly depend on disposition to trust, institution-based trust and trusting beliefs. Interestingly, such trust formation processes also seem to apply to computer systems in general [30] and to RS in particular [22].

The source of recommendation, i.e. the trustee, highly influences the acceptance of recommendations. The recommendation source, however, is not per se an automatic RS. In fact, before digitalization, recommendations were primarily provided by other humans—and often still are. The resulting two kinds of recommendation sources (i.e. human and non-human) are often termed as *personal* and *impersonal* [41, 44].

Impersonal sources that provide personalized recommendations are commonly used on contemporary online sites, but allowing other users to provide recommendations can add benefits to a service as well. Although humans have been observed to be less accurate when predicting another user's interests [25], the social cues transmitted by a personal recommendation source create social presence and can foster users' trust in a system [6, 26]. Additionally, depicting simple visual cues for trust-related attributes (e.g. expertise) of a personal recommendation source can influence trusting beliefs successfully [26].

Explaining Recommendations

Another approach to enhance trust in RS is to provide the rationale behind a recommendation in the form of textual explanations [10, 17, 46]. The literature on impact of explanations is controversial, though. While explanations have been shown to have potential for increasing transparency [42, 46], this does not necessarily improve trust in RS [8]. Yet, transparency can help users in their decision making [45] and increase user satisfaction [12]. Overall, effects of explanations seem very diverse and it can be hypothesized that this is due to different types of explanation being utilized.

One of the most common types of explanations is based on similarity between items or users and is fairly simplistic. A well-established approach, for instance, brings the recommended item into relation to those for which the user has already expressed preference. Various methods for explaining recommendations based on the computed similarity between items or users have been proposed [e.g. 2, 17, 46]. Amazon's approach of explaining recommendations based on items that were bought together constitutes another well-known example of similarity-based explanations. The effectiveness of such approaches remains questionable, though. In experiments conducted by Berkovsky et al. similarity-based RS failed to convey trusting beliefs properly [2]. Especially competence and benevolence of a recommendation source appear harder to assess based on similarity only. In line with that, Bilgic and Mooney [3] found that users in conditions with similarity-based explanations tend to overestimate the quality of recommended items, which resulted in a decrease in the perceived trustworthiness of the RS—probably due to a lower perceived competence. Yet, such explanations can result in desirable effects. Berkovsky et al. observed that similarity-based explanations can successfully increase the perceived transparency of recommendations.

However, other forms of explanations can unlock further desirable qualities. For instance, explanations indicating a high average rating of a recommendation resulted in a high perceived benevolence [2]. In the same experiment, competence was rated higher for explanations that used awards and revenue of the recommended items. Qualitative comments underlined this by assigning the latter explanation style with having the most knowledge about the item domain. In another experiment, explanations that made use of content features showed potential to increase general user satisfaction [3]. Finally, first steps have been taken for generating complex explanations based on natural language [5, 9]. Besides increasing the user satisfaction, such explanations also showed potential to be perceived as more trustworthy.

In summary, explanatory complexity spreads a continuum, ranging from rather shallow, similarity-based approaches to complex explanations that leverage natural language. Research so far gives evidence that trustworthiness of a recommendation source increases along this continuum. However, it remains underexplored which attributes of a recommendation source in particular are conveyed through such complex explanations and how. Especially, investigations are missing that shed light on how aspects such as the social presence of a recommendation source, the perceived recommendation quality and the trusting beliefs relate to each other.

3 METHOD

In order to investigate differences of personal and impersonal recommendation sources and their explanation capabilities, we conducted an online study with a between-subject design. Since we were also interested in trust dynamics over time, we conducted the experiment with two measurements over the course of two weeks. The general study setup, the

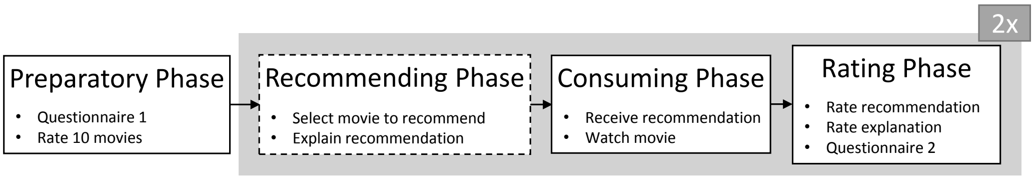


Figure 1: Study phases of the used design. Note that the phase with the dashed line (*recommending phase*) only took place in the personal condition. The phases inside the gray box were performed twice.

two conditions (personal and impersonal recommendation source), the consecutive points of measurements, as well as the tools used are described in detail below.

General Setup

As items to be recommended, we chose movies. The general rationale behind this decision was that we wanted participants to be familiar with the domain. This is a crucial point since participants had to be able to provide recommendations. A second benefit of the movie domain is the abundance of well-established datasets for automatic RS, such as the Movielens 20M rating dataset[1], which we utilized here. Since we wanted participants to be able to watch recommended items, possible recommendation candidates were restricted to those available at Amazon Prime, resulting in 393 recommendation candidates.

For the experiment we recruited 93 participants (55 female) with an average age of ($M = 25.75$, $SD = 9.00$) years. Most participants were students (68 %) or employees (24 %). A requirement for participating in the experiment was that the candidates had an Amazon Prime account so that they could actually consume recommended items. Consequently, participants were used to online streaming providers, using them on a daily (41 %) or weekly (31 %) basis. Participants were randomly assigned to conditions, resulting in sample sizes of $N = 49$ for the personal and $N = 44$ for the impersonal condition.

In a preparation step, all participants—independent of the assigned condition—were asked to rate 10 movies they already knew on a 5-point rating scale in order to elicit preferences. Afterwards, they were asked to follow the scheduled interaction cycle (see Figure 1) that was slightly varied between conditions.

Impersonal Condition

We designed the system inspired by typical online RS: after rating the items (see above), participants immediately received a recommendation. Recommendations were generated using the well-established technique of *Matrix Factorization* [24]. Specifically, we used the *Java* implementation of the *ParallelSGDFactorizer* made available by *Apache Mahout*[2]. In tandem with the recommended item, a similarity-based explanation for the recommendation was presented. Supposing that *Fight Club* was recommended and *Pulp Fiction* was highly rated by the user, the explanation had the following form:

> *Fight Club is recommended to you because it is very similar to Pulp Fiction.*

After receiving recommendation and explanation, participants were asked to watch the movie and subsequently rate movie, recommendation and explanation on a 5-point rating scale. Some days later, a new recommendation and explanation was calculated and presented. Again, participants were asked to watch the recommended movie and rate recommendation, movie and explanation afterwards.

Personal Condition

Overall, the personal condition followed the study design of the impersonal condition with one exception: All participants were assigned a *buddy*[3] and—in order to estimate preferences—were presented with the buddy's 10 rated movies. At the same interface, a searchable list of all 393 recommendation candidates, being available at Amazon Prime, was shown. Out of these candidates, participants should pick one as recommendation and compose an explanation for why they recommended it. This explanation was restricted to 255 characters in order to be comparable to the explanations from the impersonal condition in terms of length.

Instruments

We set up a website in order to deliver automatic recommendations to participants in the impersonal condition and to connect participants in the personal condition to each other. We used the same layout for both to control for confounding stimuli.

[1] https://grouplens.org/datasets/movielens/20m/; the dataset comprises 20 million ratings for 27,000 movies by 138,000 users

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

[3] In general this assignment was random but we controlled it for avoiding reciprocal relations. Participants received recommendations from a different person as they were providing recommendations to.

Several times over the course of the two weeks (see Figure 1), participants were asked to fill in questionnaires. After the first login into the system and before preference elicitation (i.e. *Preparatory Phase*), participants were asked to complete the first questionnaire on general demographics. Additionally, prior domain knowledge, the frequency of using online streaming providers and general trust in technology Knijnenburg et al. [21] were measured. Furthermore *disposition to trust* and *institution-based trust* [30, 31] were assessed. All items were measured using a 7-point Likert scale.

The second questionnaire was presented after participants had watched the first and second recommended movie respectively (*Rating Phase*). We used items from McKnight et al. [30, 31] to measure *trusting beliefs* and *trusting intentions*. For measuring *social presence* we relied on items from Gefen [13]. All items were assessed on a 7-point Likert scale. In addition, participants were asked to rate recommendations and explanations on a 5-point rating scale. We decided to incorporate post- instead of pre-consumption assessments because we assume participants can more resonably evaluate recommendations and explanations after consuming the item[4] [29].

4 RESULTS

Descriptive results of our study can be found in Table 1. They are split subject according to the experimental *condition* and the *point in time*, i.e. measurement. In order to unravel how *social presence*, *explanation quality*, and *recommendation quality* relate to each other and how they affect trust in the source of recommendation we hypothesized a structural model (see Figure 2) that we will describe in the following.

	1. Measurement				2. Measurement			
	Imp.		*Per.*		*Imp.*		*Per.*	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Trusting Beliefs	4.64	1.32	4.98	1.01	4.3	1.51	4.92	1.12
-Benevolence	4.45	1.48	5.0	1.18	4.2	1.6	4.88	1.21
-Integrity	4.53	1.53	5.1	1.11	4.27	1.55	5.08	1.15
-Competence	4.94	1.42	4.8	1.45	4.41	1.71	4.81	1.56
Trusting Intentions	4.43	1.29	4.39	1.18	4.12	1.41	4.41	1.2
-Willingness t. D.	4.56	1.44	4.27	1.47	4.14	1.59	4.36	1.6
-Follow Advice	4.75	1.54	4.71	1.48	4.28	1.58	4.65	1.5
-Give Information	3.99	1.43	4.18	1.34	3.95	1.65	4.23	1.36
Social Presence	2.38	1.43	3.8	1.67	2.39	1.5	3.73	1.61
Expl. Quality	3.23	1.25	3.76	1.12	2.84	1.33	3.88	1.18
Rec. Quality	3.91	1.21	3.67	1.18	3.51	1.33	3.92	1.08

Table 1: Mean Values and Standard Deviations for dependent variables. All variables were assessed using a 7-point Likert scale. Only explanation and recommendation quality were elicited on 5-point rating scales.

Structural Equation Modeling

Based on the number of latent constructs and observed variables we estimated the lower-bound for the sample size. With the probability level set to $\alpha = 0.05$ and a desired statistical power level of 0.8, the sample is required to be comprised of 184 observations to, at least, detect medium effects (0.3) [7, 48]. Since measurements of our experiment were taken at two points in time, we had access to 186 observations[5] in total for our analysis and are thus matching the required threshold.

We were interested in identifying whether the interaction led to differences in the assessment of trust subject to our experimental condition, i.e. a personal vs. impersonal recommendation source. *Condition* was defined as an exogenous categorical variable. We hypothesized the recommendation source not only to have an impact on trust towards the source itself (*trusting beliefs*) but also on the willingness to perform trust-related behavior (*trusting intentions*). We further assumed that this effect was mediated by systematic differences between the recommendations provided by the two sources, e.g. the nature of the explanations. Since the interaction stretched across two phases, we additionally considered whether trust would change over time. Just like *condition*, *point in time* was defined as an exogenous dummy variable. Structural equation modeling was applied to trace causal paths that lead to the development of trust or a lack thereof. For this, we utilized the *R* package *lavaan*, version 0.6-2 [39].

We conducted missing data analysis, outlier detection, a test for normality, and the selection of an appropriate estimator as preparation steps. Missing columns were observed for two participants. Little's MCAR test turned out to be non-significant ($\chi^2 = 78.01, df = 64, p = 0.11$). Therefore, we can safely assume that the data was missing completely at random and we were allowed to use maximum likelihood parameter estimation. Outlier detection based on Cook's distance revealed three rows to be outliers which were subsequently dropped leaving us with a final sample size of 184. Shapiro's test for normality indicated that several variables of interest significantly deviated from normal distributions. As a result, we conducted the analysis with an estimator that allows for robust standard errors and scaled test statistics. Together with the requirement to handle missing data, we settled with the MLR estimator [11].

Since we found no evidence that *point in time* had an influence on any constructs of interest, we decided to omit

[4] We, nonetheless, tested for possible differences between pre- and post-consumption and did not find any significant differences which is in line with [29] for the movie domain.

[5] Due to combining observations from two points in time we cannot assume mutual independence. Separate structural models for each point in time, however, revealed effects identical to the combined model. Therefore, we assume that the influence of dependence is neglegible.

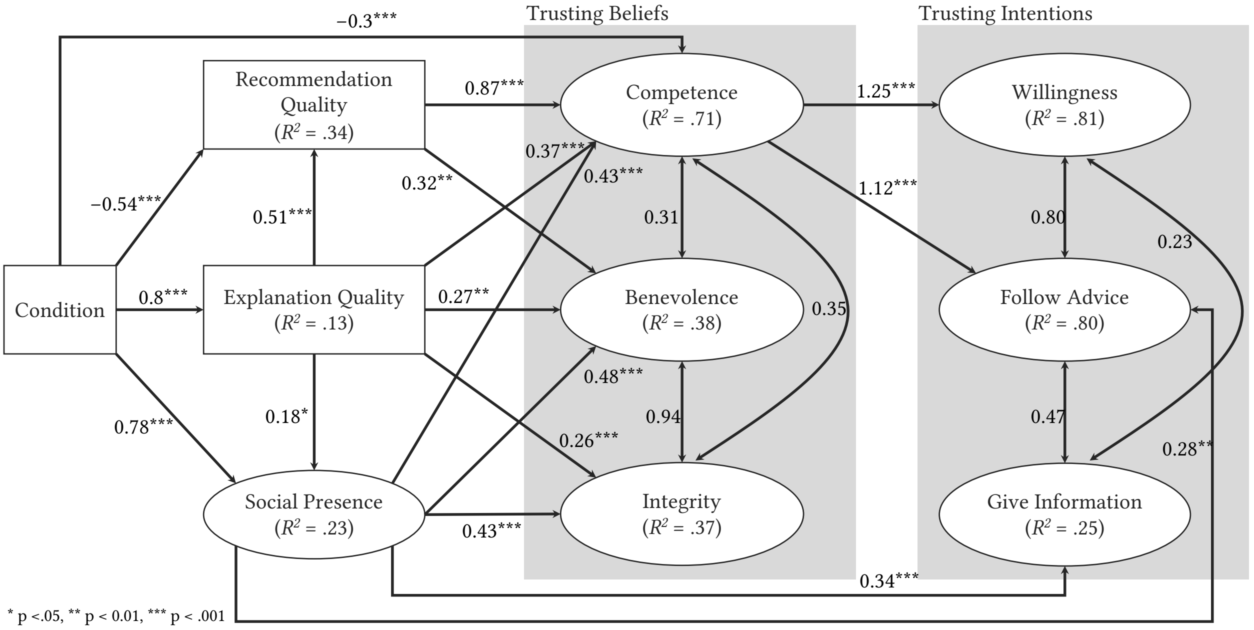

Trusting Beliefs
Trusting Intentions
−0.3***
Recommendation Quality (R^2 = .34)
0.87***
Competence (R^2 = .71)
1.25***
Willingness (R^2 = .81)
0.37***
0.43***
0.32**
1.12***
−0.54***
0.51***
0.31
0.80
0.23
Condition
0.8***
Explanation Quality (R^2 = .13)
0.27**
Benevolence (R^2 = .38)
0.35
Follow Advice (R^2 = .80)
0.48***
0.78***
0.18*
0.94
0.26***
0.47
0.28**
Social Presence (R^2 = .23)
0.43***
Integrity (R^2 = .37)
Give Information (R^2 = .25)
0.34***
* p <.05, ** p < 0.01, *** p < .001

Figure 2: Structural Equation Model comparing the influence of an algorithmic recommendation source with a human. Manifest (observed) variables are depicted as rectangles and latent (unobserved) constructs as ellipses. To prevent overloading the graph, the observed questionnaire items corresponding to latent variables are omitted. The edges show standardized parameter weights and the amount of explained variance for endogenous variables is displayed inside the nodes.

it. The remainder of our hypothesized model appears to be a good fit for the data ($CFI = .970$, $TLI = 0.963$, $RMSEA = 0.052$). For the sake of clarity, we will report significant direct effects successively from left to right. Along these paths we will trace back mediated influences from *condition* on the endogenous variables.

Direct Effects & Mediation via Explanation Quality. The positive direct effect from *condition* onto *explanation quality* (see Figure 2) suggests that the explanations formulated by human buddies attain higher quality than the generic similarity-based ones. *Explanation quality* acts as a mediator between *condition* and *recommendation quality* as well as between *condition* and *social presence*.

While *condition* has a negative direct effect on *recommendation quality*, suggesting that human buddies provide recommendations of lower quality, the mediation [*condition* → *explanation quality* → *recommendation quality*] yields a competing impact of .44 ($p < .001$). Hence, although recommended movies from a personal source are perceived as worse if examined in isolation, this effect is antagonized by the significant positive influence exhibited by the explanations provided. When put together, both effects cancel each other resulting in a total effect of 0.02 ($p = .919$).

Concerning *social presence*, the direct as well as the indirect effect [*condition* → *explanation quality* → *social presence*] (standardized coefficient = 0.12, $p = 0.03$) assume positive polarity. The combined total effect is 0.9 ($p < .001$) indicating that having a personal recommendation source is related to higher levels of *social presence.*

Direct & Indirect Effects on Trusting Beliefs. We can observe four direct effects on *competence*: The negative impact from *condition* suggests that, per se, the human buddy is perceived as less competent. The remaining three effects from *recommendation quality*, *explanation quality*, and *social presence* are all positive with, according to its parameter weight, *recommendation quality* having the strongest influence.

Both *recommendation quality* and *social presence* thereby become mediators themselves carrying some of the explanatory power of *condition* and *explanation quality.* For instance, the path [*condition* → *social presence* → *competence*] yields an indirect effect of 0.21 ($p < .001$). Please note that we now also have to consider paths with two mediators such as [*condition* → *explanation quality* → *recommendation quality* → *competence*] with an effect of 0.21 ($p = .001$). Combining all these effects leads to a non-significant total effect of .11 ($p = .475$) from *condition* on *competence.*

There exist similar causal patterns for *benevolence* except for the insignificant direct effect from *condition.* As a result, all explanatory power can be distributed to the mediators. The total effect of 0.453 ($p = .002$) tells us that personal recommenders cause participants to develop higher levels of benevolence via better *explanations* and increased *social presence* despite the negative impact of lower *recommendation quality.*

The strongly positive total effect from *condition* on *integrity* with 0.51 ($p < .001$) suggests that personal recommenders appear more honest and genuine than their automated counterparts. Again, we cannot identify a significant direct influence from *condition* such that all causal effects can be explained by means of mediators. While the indirect paths via *explanation quality* and *social presence* depict similar patterns as the ones discussed for *competence* previously, the effect from *recommendation quality* turned out to be non-significant.

Direct & Indirect Effects on Trusting Intentions. The causal influence on *willingness to depend* can exclusively be reduced to *competence* as it is the only significant predictor. Therefore, it is sufficient to only analyze the paths to that point as the implications are equivalent. By combining the previous non-significant effect from *condition* on *competence* with the direct impact from the latter, we obtain a total effect of 0.26 ($p = 0.161$).

Since *competence* is also an influencing factor for *follow advice*, the same relationships as for *willingness to depend* are of importance again. Additionally to the effects via the route [*condition* → *(mediators)* → *competence*], there is a significant direct influence from *social presence* of 0.39 ($p < .001$) this time. Elevated *social presence* therefore leads to a greater tendency to follow the recommender's advice. Put together, the total effect is 0.41 ($p = .028$).

Give information is completely independent of any paths that tackle recommendations or even the recommendation source. Exclusively by an increased *social presence* is it possible to predict a higher probability of a person sharing information (standardized coefficient = 0.39, $p = .005$).

Counterfactual Analysis

The structural model described in the previous section has already provided some insights into causal effects exhibited by the exogenous exposure variable *condition.* By decomposing its total effect into direct and indirect parts, we have exposed *explanation quality* as the pivotal discriminating factor between personal and impersonal recommendation sources. Due to the generic nature of the explanations generated by the RS, its trustworthiness and *recommendation quality* as well as perceived *social presence* were obviously confined.

On the basis of these findings, we can now hypothesize that RS performance is likely to be substantially improved if better explanations could be provided. The counterfactual mediation framework allows us to investigate questions about such hypothetical situations with outcomes we cannot observe in reality. Specifically, counterfactual analysis lets us express the potential change induced by the *condition* when keeping *explanation quality* fixed at the value that had naturally been observed. In other words, we can estimate the degree to which, for instance, *recommendation quality* would change if the RS was capable of generating explanations of the same quality as humans.

We can achieve this in terms of composite or nested counterfactuals. Let $Y_i(x, M(x))$ be the outcome for individual i when exposed to *condition* x under consideration of the mediator's M influence. For binary exposures, the composite counterfactual is then the outcome for *condition* x subject to the intermediate outcome for the alternate exposure level x^*, i.e. $M(x^*)$. Generalizing to population level is done by taking the expected value which yields the mediation formula [35]:

$$\mathrm{E}\{Y(x, M(x^*))\} = \sum_m \mathrm{E}(Y(x, m))\mathrm{Pr}(m|x^*, C), \tag{1}$$

where C is a set of confounding variables. Since we are interested in the expected improvement over actually observed values for the mediator, $\mathrm{E}\{Y(x, M(x))\}$, we need to calculate the unit effect **UE** of M on Y given X:

$$\mathbf{UE} = \mathrm{E}\{Y(x, M(x^*))\} - \mathrm{E}\{Y(x, M(x))\} \tag{2}$$

We calculated a mediation model with the outcome set to *recommendation quality* in order to emphasize the importance of explanations to support the main goal of RS, i.e. generating good recommendations. Therefore we set $Y = \textit{recommendation quality}, M = \textit{explanations quality}, X = \textit{condition}$. Based on the results of the structural model and further investigations, no confounding variables could be identified. The resulting unit effect is:

$$\mathbf{UE}(\text{recommendation quality}) = 4.11 - 3.63 = 0.48 \tag{3}$$

Altering *condition* from *personal* to *impersonal* while maintaining *explanation quality* therefore increases the expected assessment of *recommendation quality* from 3.63 to 4.11 which corresponds to an improvement of 13 %.

Qualitative Analysis of Explanations

In order to get a better understanding of how participants composed explanations, we provide some examples (see Table 2). Examining those examples more closely shows that explanations vary from very sophisticated statements (e.g. p294) to shallow comments (e.g. p273). Some also use similarities (e.g. p435) or express uncertainty (e.g. p369). Others address general quality of the recommended movie (e.g. p414) or try to be convincing and flattering (e.g. p427). Overall, generated explanations used a similarity relation to the rated movies of the recommendation receiver in 37 %.

Participant	Sample Explanations
p269	"*Based on the rated movies of the buddy I don't know what he likes or dislikes. Hence i chose an entertaining over the top action movie, that is diverting for the short time of the movie.*"
p270	"*I chose the film because I saw it myself and was excited about it. My buddy and I seem to have a similar taste. Besides I wanted to pick a movie in the genre of fantasy/science fiction, based on the rated movies of the buddy.*"
p273	"*Fantasy movie, action*"
p294	"*A classic and atmospheric story, where a noble-minded hero fights an epic battle against the evil (as in most of my buddy's highly rated movies).*"
p307	"*Once is a low budget movie, that has a lot to offer musically. My buddy seems to like films that are emotional and do have melancholic soundtracks. Therefore i chose this nonfamous movie.*"
p369	"*I find it difficult to find a matching movie since the genres of your rated movies are quite different. In addition I do not know most of them. I recommend Disturbia as it mixes action and thriller elements and hope that matches your taste. :)*"
p414	"*A thrilling movie with a tangled plot of hunter and hunted, awesome cast and a whole lot of action.*"
p421	"*I think you don't like romantic comedy or extreme horror movies. As a result I picked this movie. It contains action not too much and a good story that concludes with the movie. Have fun!*"
p427	"*Memonto is very thrilling to watch and contains a whole bunch of light bulb moments. I think this movie is very sophisticated and nothing for bores—thus the perfect movie for guys who like profound stories, like you ;)*"
p435	"*Since my Buddy rated Forrest Gump highly, I guess he/she will like this touching movie with Tom Hanks as well.*"
p445	"*Because it's a good movie*"

Table 2: Some of the explanations created by participants in our experiment (carefully translated to English).

Taking into consideration the language style, 16 % of explanations addressed the buddy directly, 44 % used the third person and 38 % were formulated in a neutral manner. Smileys or other kinds of emoticons were only used scarcely (in 10 % of the explanations). 18 % of the explanations expressed a high certainty regarding the recommended movie, whereas in 4 % of cases it was explicitly stated that the participants were not sure about the recommendation. On average participants used $M = 23.23$ ($SD = 10, 71$) words for their explanations[6].

Explorative Inspection of Temporal Effects

Although the structural model revealed no significant differences for the point in time, we were still interested in explorative investigation. Overall, the reported values in Table 1 are homogeneous within conditions. This is underlined by statistical comparisons: When comparing results between points of measurements, there were no significant differences—neither in assessed quality of recommendations and explanations nor regarding trust in the source of recommendation. Only for the impersonal condition, statistical significant differences are found. Concretely, with ($t(43) = 1.989, p = .053$) *trusting beliefs* were higher at the first point of measurement. This is also true for its subconstruct *competence* ($t(43) = 1, 973$, $p = .055$). Although values for *benevolence* and *integrity* seem to decrease slightly over time, this was not significant. Similar observations can be found regarding *trusting intentions*. Within the impersonal condition, we also found a marginal significant difference here ($t(43) = 1, 984$, $p = .054$). Again, the values at the first point of measurement were slightly higher. This also holds for the subconstruct *follow advice* ($t(43) = 2.126$, $p = .039$). Values of *willingness to depend* were not significant, but seem to slightly decrease, whereas the intention to *give information* nearly remains stable over time.

Summary of Findings

The main focus of the statistical analysis presented was the investigation of causal paths along a structural model (Figure 2) that lead from the effects of our experimental *condition* to trust-related constructs. Our results suggest that the higher-quality explanations provided by participants had an overall positive effect on their buddies' *trusting beliefs* and *trusting intentions*, despite the lower *recommendation quality*. We discuss the implications of our findings in the next section.

5 DISCUSSION

Close inspection of the relations discussed in the previous section hint at a pivotal role of *explanation quality*. *Recommendation quality*, *social presence* and the trusting beliefs *competence*, *benevolence* and *integrity* were all significantly and directly affected by the quality of explanations.

Recommendation and Explanation Quality

By distinguishing between direct and indirect influences, we were able to detect systematic effects that would otherwise have been obstructed. Concretely, no differences could be found descriptively between the two conditions for perceived *recommendation quality* (see Section 4). However, by taking into account the mediating function of explanations, a negative direct effect became evident. That is, if we look at the chosen items in isolation and control for any influence explanations might have, human recommendations were less likely to conform to the receiver's preferences. This finding is in line with previous research[25]: Humans tend to listen

[6] Explanations were restricted to a maximum of 250 characters.

to their “gut feelings” and rely on vague emphatic estimations, whereas the RS, due to its statistical nature, has access to a vast factual basis from which to derive its decision.

However, the parameter weights on the indirect path [*condition → explanation quality → recommendation quality*] cause the total effect to become insignificant. Our deductions are twofold: First, a good explanation can, at least to a certain degree, make up for a poor recommendation. Second, humans compose explanations of significantly superior quality[7].

We originally solicited *movie quality* besides *recommendation quality* but discarded it. While in some cases there surely is a difference (imagine recommending a movie the user already knows and likes: even though the item is liked, the recommendation would not be considered very helpful), it seems that participants in our experiment could not draw a mental line between these concepts. When replacing *recommendation quality* with *movie rating* inside the SEM, effects stay identical. This should not be the case had the movie been rated solely on watching experience. Especially *explanation quality* would not have influenced subjective *movie quality*.

People dislike the explanations generated by the RS because they are, in essence, a verbalization of the similarity relation between a previously rated movie and the recommendation. Without any further context given, they appear arbitrary to users. In our experiment the RS did not disclose its decision criteria, thus making it difficult for participants to understand the foundation for the similarity estimation. Moreover, the system did not explain why a particular rated item was chosen as the basis for an explanation and not any other.

On the other hand, humans conveyed their explanations in an argumentative manner that resembles very closely the process of how people exchange information in reality. Overall, they gave more nuanced explanations by justifying their choice and contextualizing the recommendation with respect to a plurality of dimensions. Interestingly, they often also used similarity to rated movies, revealing that this style is per se suitable for explanation purpose. Yet, humans often combined several explanation styles, e.g. by summarizing content commonalities in rated movies and bring them into similarity-context with the recommendation (e.g. see p307 of Table 2). We thus believe that combinations of different explanation styles lead to explanations with a higher perceived value, which is backed up by prior findings [5].

RS in general should be equipped with more sophisticated means of explaining their decisions. Counterfactual inference give us concrete hints about the effect size we can expect: While maintaining the same algorithmic accuracy and only by adopting to a human-like rather than a similarity-based explanation style, the quality of recommendations would be improved, on average, by around 0.5 points on the rating scale. Expected improvements over RS that do not provide any explanations at all—which are still very common—would be even greater.

Trusting Beliefs and Trusting Intentions

Beyond improving the quality of recommendations, our experiment shows that good explanations can also increase the trust in their authors as expressed by the significant effects on all subconstructs of *trusting beliefs*. It is safe to assume that individuals who can articulate profoundly how they chose a movie as recommendation, e.g. by contextualizing their choice, will be considered competent advisors. Moreover, integrating direct speech and other subtleties of human language into the explanation text may trigger associations of benevolence and integrity. These competences, which humans learn naturally through socialization, are typically not reflected in RS explanations. Lower values in *explanation quality* and therefore trustworthiness are possible consequences. This assumption is underlined, although not with statistical significance, by the fact that we observed diminishing trust over time in the RS that was not traceable for humans. After a high initial trust, which is not uncommon when establishing new relations [31], users were supposedly disappointed by the explanatory capabilities of the system. The resulting asymmetry of information and unfulfilled expectations probably led to the observed decrease in trust. As a consequence of these factors, we suggest developing systems for incorporating explanatory components in a manner that resembles more closely the way in which humans exchange information.

Apart from that, we found some interesting relations regarding the subconstructs of trusting beliefs that we shortly want to discuss: First, there was a direct (negative) effect from *condition* on *competence*. That is, a priori RS are perceived as more competent, likely because of a dispositional attitude people seem to have. Second, although the total effect from *condition* on *benevolence* indicates that humans are assessed as being more benevolent than a machine, it is still surprising that we could not identify a predisposition—expressed by a significant direct effect—in favor of humans that transcends the indirect influences. Third, *recommendation quality* seems to have no effect on *integrity*. This can easily be explained against the background that assessing someone as upright is rarely connected with the perception of how good they are at a particular task.

The prediction of *trusting intentions*—and thus trust-related behavior—on the basis of the degree of trust into a source of recommendations is, in contrast to prior research [e.g.

[7] Please note that the R^2 value in explanation quality is rather low at 13%. We account this to the fact that our binary exposure variable obviously cannot explain variance that occurs within *conditions* but only between.

30], only possible via *competence*. We assume that *benevolence* and *integrity* were not conveyed sufficiently in our experimental setup. Moreover, the movie domain may not create the necessity of such traits in order to follow an advice. Thus, we believe that with more information about the recommendation source available, further communication possibilities, and an item domain in which such traits are more important (e.g. real estate business), benevolence and integrity would become more influential. Interestingly, the effect on *trusting intentions* outgoing from *social presence* does not get completely mediated through *trusting beliefs*. *Social presence* directly influences the tendencies to *follow advice* and to *give information*. Certain undetected social cues seem to have been present during interaction, distinct from the recommendation source, that facilitate such social behavior.

Social presence itself is partially affected by *explanation quality*. We can observe at least a small effect that indicates that better explanations increase social awareness. The major portion of explained variance, however, originates in the differences between the two *conditions*. The knowledge about whether the interlocutor is human or not significantly influences one's perception of being in a social situation. This effect seems also to occur through the restricted information channel determined by the recommendation platform which corresponds to prior research [6, 26]. One important factor for the observed perceived *social presence* is probably also the conceptualization of the user study as a reciprocal act between humans. Users in the human condition were not only receivers of recommendations but also producers. This will likely lead to elevated feelings of social exchange and probably also to situational sympathies.

Finally, there are some limitations to our study to consider. We are aware of possible biases in our sample since a requirement for taking part in our study was to have a streaming account. For this reason, we believe that participants were somewhat technically skilled and probably less picky about recommendations and their explanations. Additionally, only a single item was recommended for each point in time, which is not a typical RS situation. This decision was made because we wanted to isolate the situation from as many stimuli as possible. If more than one recommendation would have been shown, other aspects, such as diversity, may have influenced perceived quality of recommendations. However, since this was the case in both conditions, possible biases (e.g. on perceived *recommendation quality*) can be neglected.

6 CONCLUSIONS AND FUTURE WORK

We have provided a detailed analysis of the causal effects that determine the outcome of trust in personal vs. impersonal recommendation sources. We laid particular focus on exposing systematic effects that can be causally ascribed to the fundamental differences between personally composed and automatically generated explanations. Structural equation modeling offered us the tools to uncover subtle cause-effect relationships. By tracing back indirect influences over elongated paths we could identify the relative impact of *recommendation quality*, *social presence*, and especially *explanation quality* on trust. Thereby, our structural model provides an indication of general mechanisms relevant for generating good recommendations that could not have been derived with correlative studies. Counterfactuals helped us answer questions about hypothetical situations in which RS are able to generate human-like explanations. Unit effect values indicate that being capable of doing so will likely turn out to have a significant impact on the perceived quality of recommendations. The impersonal nature of automated RS can, at least to some degree, be overcome by approaching an explanation style that humans tend to employ in everyday interaction.

On the basis of these results, we conclude that the positive impact of adequate explanations is considerably underestimated and receives too little attention in research and—even more decisively—in industry. If we look at contemporary explanations on online platforms, they are, if anything, a subordinate component, be it in Netflix, Spotify or YouTube. We argue for a more prominent role of explanations in RS—especially due to the mediating effects of explanation quality: While automated RS seem to generate recommendations of superior quality, this benefit is countered by the quality of human explanations to the degree of complete equalization. In other words, the tremendous accuracy of recommending algorithms, emerging from decades of research in that area, remains next to meritless, when RS fail to convey rationales behind their recommendations.

Finally, considering the trend of incorporating more and more natural language into human–computer interaction (e.g. personal voice agents such as *Siri* or Amazon's *Echo*), in future work we will aim at analyzing human-generated explanations in more detail to derive insights into features used and their impact on trust. We also plan to utilize more sophisticated explanations in our experimental setting and intend to take conversational explanation patterns into consideration, enabling RS to answer on specific questions about recommendations.

REFERENCES

[1] James Bennett and Stan Lanning. 2007. The netflix prize. In *Proceedings of KDD Cup and Workshop (KDDCup '07)*, Vol. 2007. San Jose, CA, 3–6.

[2] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, 287–300. https://doi.org/10.1145/3025171.3025209

[3] M. Bilgic and Raymond J. Mooney. 2005. Explaining recommendations: satisfaction vs. promotion. In *Proceedings of Beyond Personalization Workshop, IUI.*

[4] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for Recommender Systems: The Past, the Present and the Future. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).* ACM, New York, NY, 123–126. https://doi.org/10.1145/2959100.2959158

[5] Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).* ACM, New York, NY, 175–182. https://doi.org/10.1145/2959100.2959153

[6] Jaewon Choi, Hong Joo Lee, and Yong Cheol Kim. 2009. The influence of social presence on evaluating personalized recommender systems. In *Pacific Asia Conference on Information Systems (PACIS).*

[7] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455–496. https://doi.org/10.1007/s11257-008-9051-3

[9] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2018. Explaining Recommendations by Means of User Reviews. In *Proceedings of the 1st Workshop on Explainable Smart Systems (ExSS '18).* http://ceur-ws.org/Vol-2068/exss8.pdf

[10] A. Felfernig and B. Gula. 2006. An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE '06).* 159–169. https://doi.org/10.1109/CEC-EEE.2006.14

[11] David A Freedman. 2006. On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician* 60, 4 (2006), 299–302. https://doi.org/10.1198/000313006X152207

[12] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382. https://doi.org/10.1016/j.ijhcs.2013.12.007

[13] David Gefen. 1997. *Building Users' Trust in Freeware Providers and the Effects of This Trust on Users' Perceptions of Usefulness, Ease of Use and Intended Use of Freeware.* Ph.D. Dissertation. Atlanta, GA.

[14] Carlos A. Gomez-Uribe and 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

[15] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 265–308.

[16] Ido Guy. 2015. Social Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 511–543.

[17] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00).* ACM, New York, NY, 241–250.

[18] Miguel Angel Hernán. 2004. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health* 58, 4 (2004), 265–271. https://doi.org/10.1136/jech.2002.006361

[19] Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. Human Decision Making and Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 611–648.

[20] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders–Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10:1–10:46. https://doi.org/10.1145/3001837

[21] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[22] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The Effects of Personalizaion and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4 (2006), 941–960. https://doi.org/10.2307/25148760

[23] Joseph A. Konstan and 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

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

[25] Vinod Krishnan, Pradeep Kumar Narayanashetty, Mukesh Nathan, Richard T. Davies, and Joseph A. Konstan. 2008. Who Predicts Better?: Results from an Online Study Comparing Humans and an Online Recommender System. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08).* ACM, New York, NY, 211–218.

[26] Johannes Kunkel, Tim Donkers, Catalin-Mihai Barbu, and Jürgen Ziegler. 2018. Trust-Related Effects of Expertise and Similarity Cues in Human-Generated Recommendations. In *Companion Proceedings of the 23rd International on Intelligent User Interfaces: 2nd Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE).* http://ceur-ws.org/Vol-2068/humanize5.pdf

[27] Johannes Kunkel, Benedikt Loepp, and 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 Companion (IUI '17).* ACM, New York, NY, 3–15. https://doi.org/10.1145/3025171.3025189

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[29] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2018. Impact of Item Consumption on Assessment of Recommendations in User Studies. In *Proceedings of the 12th ACM Conference on Recommender Systems.* ACM, New York, NY, 49âĂŞ53. https://doi.org/10.1145/3240323.3240375

[30] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[31] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. 1998. Initial Trust Formation in New Organizational Relationships. *Academy of Management Review* 23, 3 (1998), 473–490. https://doi.org/10.5465/amr.1998.926622

[32] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06).* ACM, New York, NY, 1097–1101. https://doi.org/10.1145/1125451.1125659

[33] Bengt Muthén. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 1 (1984), 115–132. https://doi.org/10.1007/BF02294210

[34] Umberto Panniello, Michele Gorgoglione, and Alexander Tuzhilin. 2015. Research Note—In CARSs We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems. *Information Systems Research* 27, 1 (2015), 182–196. https://doi.org/10.1287/isre.2015.0610

[35] Judea Pearl. 2012. The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models. *Causality: Statistical Perspectives and Applications* (2012), 151–179. https://doi.org/10.1002/9781119945710.ch12

[36] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556. https://doi.org/10.1016/j.knosys.2007.04.004

[37] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355. https://doi.org/10.1007/s11257-011-9115-7

[38] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 1–34.

[39] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. http://www.jstatsoft.org/v48/i02/

[40] Donald B Rubin. 1990. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference* 25, 3 (1990), 279–292. https://doi.org/10.1016/0378-3758(90)90077-8

[41] Sylvain Senecal and Jacques Nantel. 2004. The influence of online product recommendations on consumers' online choices. *Journal of Retailing* 80, 2 (2004), 159–169. https://doi.org/10.1016/j.jretai.2004.04.001

[42] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, 830–831. https://doi.org/10.1145/506443.506619

[43] B. Smith and G. 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

[44] Donnavieve Smith, Satya Menon, and K. Sivakumar. 2005. Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing* 19, 3 (2005), 15–37. https://doi.org/10.1002/dir.20041

[45] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 399–439. https://doi.org/10.1007/s11257-011-9117-5

[46] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 353–382.

[47] Patricia Victor, Martine de Cock, and Chris Cornelis. 2011. Trust and Recommendations. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, Boston, MA, 645–675. https://doi.org/10.1007/978-0-387-85820-3_20

[48] J Christopher Westland. 2010. Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications* 9, 6 (2010), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

[49] Sherrie Xiao and Izak Benbasat. 2003. The Formation of Trust and Distrust in Recommendation Agents in Repeated Interactions: A Process-tracing Analysis. In *Proceedings of the 5th International Conference on Electronic Commerce (ICEC '03)*. ACM, New York, NY, 287–293. https://doi.org/10.1145/948005.948043

[50] Jingjing Zhang and Shawn P. Curley. 2018. Exploring Explanation Effects on Consumers' Trust in Online Recommender Agents. *International Journal of Human–Computer Interaction* 34, 5 (2018), 421–432. https://doi.org/10.1080/10447318.2017.1357904

The following article is reused from:

Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 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, pages 3–15. ACM, 2017. ISBN 978-1-4503-4893-5. doi: 10.1145/3025171.3025189

A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering

Johannes Kunkel
University of Duisburg-Essen
Duisburg, Germany
johannes.kunkel@uni-due.de

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

While conventional Recommender Systems perform well in automatically generating personalized suggestions, it is often difficult for users to understand why certain items are recommended and which parts of the item space are covered by the recommendations. Also, the available means to influence the process of generating results are usually very limited. To alleviate these problems, we suggest a 3D map-based visualization of the entire item space in which we position and present sample items along with recommendations. The map is produced by mapping latent factors obtained from Collaborative Filtering data onto a 2D surface through Multidimensional Scaling. Then, areas that contain items relevant with respect to the current user's preferences are shown as elevations on the map, areas of low interest as valleys. In addition to the presentation of his or her preferences, the user may interactively manipulate the underlying profile by raising or lowering parts of the landscape, also at cold-start. Each change may lead to an immediate update of the recommendations. Using a demonstrator, we conducted a user study that, among others, yielded promising results regarding the usefulness of our approach.

Author Keywords

Recommender Systems; Interactive Recommending; Matrix Factorization; User Interfaces; User Profiles; User Experience; 3D Visualizations

ACM Classification Keywords

H.3.3. Information Storage and Retrieval: Information Search and Retrieval—*information filtering*; H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces—*evaluation/methodology, graphical user interfaces (GUI), user-centered design*; I.3.8. Computer Graphics: Applications

INTRODUCTION

Recommender Systems (RS) have become a widely adopted means to tackle the problem of information overload users are often confronted with, for instance, on e-commerce websites, in social networks, on hotel booking portals or in online movie stores [52]. To present users with items that meet their interests, different approaches have emerged. *Collaborative Filtering* (CF), the overall most popular recommendation technique, solely relies on user feedback elicited by asking users explicitly to rate items or by implicitly tracking their interaction with the systems [26, 36]. *Matrix Factorization* (MF) represents the most common model-based CF approach, which generally performs best in terms of objective accuracy while being highly efficient [37]. By statistically analyzing existing rating data, latent factors are inferred which in the following can be used to predict a user's ratings for yet unseen items.

For a long time, RS research has solely been focused on issues related to such algorithms, in particular their accuracy and performance. Only recently, it became more and more accepted that user-oriented aspects such as system transparency or the degree of control users are able to exert over the recommendation process considerably contribute to actual user satisfaction [66, 35, 51, 32]. For instance, users may be reluctant to accept recommendations because they do not understand why certain items are recommended [60], which consequently reduces the system's trustworthiness [66, 51]. The widely used presentation of results in form of ranked lists is not very supportive in this regard, since they usually convey only little information about the recommender's internal rationale [45, 28]. Several approaches exist to increase transparency, e.g. through explanations [24, 63, 60]. However, this typically requires additional content data and is particularly difficult when using model-based CF [37, 13]. Moreover, when presenting just top-n recommendations, users are unable to get an overview of the naturally large item space and cannot adequately assess item coverage, i.e. how shown items relate to remaining non-recommended ones. Becoming aware of alternatives and different, possible diverse areas of potential interest is thus rather difficult [46], and increases the risk of users being trapped in "filter bubbles" [47]. In addition, it often remains unclear how expressed preferences actually correspond to the system's representation of the user, i.e. the user model, and how manipulating the preference profile, e.g. by providing further ratings, affects the results.

From an algorithmic perspective, it becomes increasingly difficult to further improve how recommendations are tailored towards the user's actual needs. By providing a higher degree of control over the recommendation process, interactive RS aim at alleviating this problem in various ways [40, 21]. However, in today's RS, results are mostly adapted automatically based on implicit feedback, e.g. viewing or buying actions. To actively influence recommendations, the user's only means is usually to rate single items, either at cold-start or later in the

IUI 2017, March 13 - 16, 2017, Limassol, Cyprus

ISBN 978-1-4503-4348-0/17/03. . . $15.00
DOI: http://dx.doi.org/10.1145/3025171.3025189

process. The implicit way to elicit preferences is again prone to be intransparent while the explicit rating of items requires considerable effort on part of the user before receiving fitting recommendations [19, 58, 11]. In addition, ratings tend to be inaccurate [3] and users are shown to often prefer other means than ratings. For instance, comparing items [42] or stating interests on a rather coarse level by selecting and weighting tags [12] can be of benefit—especially for users entering a system [58]. When no or only little information is available for a new user, conventional CF suffers from the well-known cold-start problem, and thus cannot generate accurate results due to lack of data. This may also be the case when a user does not want a profile representing his or her preferences to be persisted, e.g. due to privacy concerns. Even with an existing profile, it can be difficult to recommend items matching the current user's situation since profiles usually describe long-term interests and do not necessarily need to belong to the same person, e.g. when shared between family members.

In this paper[1], we consequently propose an interactive recommending approach, thereby seeking to answer the following research questions:

RQ1: How can the item space in CF be visualized and sampled in a comprehensible manner?
RQ2: How can areas of preferred items be effectively highlighted within this visualization?
RQ3: How can this visualization be used to allow the user to interactively manipulate his or her preference profile, also in cold-start situations?

First, to visualize the item space, we apply model-based CF due to its proven precision and efficiency. In particular, we use a standard MF algorithm, but map the resulting high-dimensional latent factor model onto a two-dimensional surface in which all items are positioned with respect to their similarities. For this purpose, we use *Multidimensional Scaling* (MDS) [6, 28]. By displaying popular and representative sample items, we are then able to provide the user with a comprehensible presentation of the item space only by means of ordinary rating data, i.e. without requiring any other item-related content. Second, we extend the resulting map to also show the preferences of the current user. To reveal areas of interest, and in particular to highlight the items automatically recommended by the system and how they relate to the typically very large rest of the item space, we additionally exploit the third dimension. Therefore, we use the MF predictions for the current user and all items in order to form a landscape where elevations represent areas with high estimated ratings while valleys indicate lower relevance. Finally, this 3D visualization of item space and user preferences allows us to let the user influence the underlying profile that serves to generate recommendations. The user can alter the landscape by creating or reshaping hills and valleys, and thus establish a preference profile in cold-start situations or manipulate an existing one. All changes may immediately be reflected in the recommendations. Since preferences are expressed with respect to entire item regions rather than individual items, this reduces interaction effort and is independent of knowing and rating particular items, which is especially of value when the search goal is vague or the domain unknown.

The remainder of this paper is organized as follows: First, we discuss work related to visualizations in RS as well as interactive recommending approaches. Next, we describe our method and a prototype system we implemented to demonstrate our approach. Then, we present a user study we conducted to evaluate our method. Finally, we conclude by discussing results and providing an outlook on future work.

VISUALIZATIONS IN RECOMMENDER SYSTEMS RESEARCH AND INTERACTIVE APPROACHES

Increasing the transparency of RS is known to, among others, improve perceived recommendation quality, leading to higher acceptance and more trust in the systems [66, 51, 60]. However, today's automated recommenders often hinder users to understand *how* a system generates recommendations and *why* it recommends certain items [57, 60]. One popular approach to alleviate this problem is to display textual explanations for recommended items [63, 60]. Thus, depending on recommendation algorithm as well as type and amount of available information, recommendations can be explained in several ways. For instance, one can use item attributes and match them with user preferences [63], albeit this requires availability of content data. Social explanations have been shown to be particularly promising in terms of persuasiveness, but are less informative than other variants [55]. When using CF, a very prominent yet simple example is the one of explaining item-based methods, e.g. used by Amazon ("Customers who bought this item also bought. . . "). Nevertheless, while there exist many early attempts to explain the output of CF algorithms in general [24], especially for model-based approaches such as MF it is still very hard to improve their transparency through explanations [37]. Exceptions such as [13] usually also require additional content information.

Apart from textual explanations, the range of attempts to increase transparency of RS also includes use of visualizations. Rather simple auxiliary graphics depict, for instance, which criteria selected by a user could be fulfilled [41] or which algorithm was responsible for a recommendation in a hybrid setting [49]. But, also more complex visualizations such as flow charts [27], Venn diagrams [49] or even graph-based representations [62] have already been discussed. On a different level, other approaches visualize the user model in order to improve the system's general transparency and the user's understanding of how his or her preferences are represented within the system. Leveraging *Information Visualization* techniques [30, 23], successful examples comprise focus-and-context lists [61], radial displays [4] or icon-based avatars [5].

Particularly in *Information Retrieval*, a considerable number of methods exist for visualizing large datasets such as document collections [22, 1]. Map visualizations, for example, have been shown to be a promising means for facilitating browsing and searching in large collections. Adopted in RS

[1] This paper is a translated and extended version of our previous work published in German [38]. We now describe the method in more detail, present a more developed version of our demonstration system, and report further results from the user study.

research, maps may also be useful for visualizing the space of available items as well as the user model [44, 65, 17], possibly making the recommendation process more transparent as well as increasing user engagement when interacting with the RS [17, 15]. Assuming a user's preference profile is represented by some high-dimensional vector, it can be projected onto a 2D representation of the item space as a geographical point where items that receive a high predicted rating appear close to this point [31, 15, 44]. Thus, the relation between how the user is modeled and which items are recommended becomes intuitively understandable [31]. However, this kind of maps highly depends on the particular user: They cannot be generated without sufficient information about that user, e.g. at cold-start, and items are arranged differently for each user. Moreover, these approaches usually require a version of the underlying algorithm where the rating prediction is specially geared to create such maps, e.g. by using Euclidean MF [31, 44]. The map visualization of *TVLand* [17], in contrast, is independent of a particular user and his or her estimated ratings. Here, similarities between items are used to create a global representation of the item space. Nonetheless, areas of interest that include the recommendations can still be highlighted by color, similar to a heat map. Consequently, users can see how their preferences expressed through ratings result in the areas and items the system actually suggests. In addition, users are better able to grasp how recommended items are positioned within the rest of the typically large item space, thus allowing them to keep an overview and to become aware of possible alternatives. In conventional RS, this is often difficult due to their use of lists to present recommendations.

Approaches that visualize only a part of the item space, for instance, the region close to the user's position containing the recommended items (as in [31, 44]), may also be prone to this problem. To mitigate the risk of users thus being stuck in a "filter bubble" [47], some visualizations specifically aim at presenting diverse recommendations [48, 65]. Only few exceptions such as the aforementioned *TVLand* [17] visualize the item space as a whole, and at the same time also indicate areas of potential interest. Overall, effectively supporting users through visualizations in RS is still an under-explored field of research, mostly limited to the purpose of explaining recommendations or supporting item space exploration. In addition, possibilities to interact with the visualizations almost never go beyond the means provided in conventional RS.

In CF, preferences are usually elicited via implicit or explicit feedback [26]. However, providing explicit feedback, typically by rating single items, is a tedious task for users that is often decoupled from the actual recommendation process. At the same time it constitutes a very limited means for expressing actual user needs. Thus, this kind of feedback is rather sparsely available [26]. Users who enter a system for the first time have to rate a certain number of items before a CF algorithm can provide them with proper results [11]. To counter this, efforts have been made to keep the number of items to be rated as small as possible [43], to reward users for every rating provided [16], or to seek for alternatives, e.g. comparing items instead of rating them [42]. Also, algorithmic solutions have been suggested with the goal of asking users to rate only the most informative items, e.g. via active learning [14]. However, expressing initial preferences as well as altering an existing profile is nearly impossible in a controlled and transparent manner when the user's only way to influence the recommendation process is to (re-)rate single items.

Therefore, interactive recommending approaches have been proposed that increase user control over the recommendation process. It has been well established that users are generally more satisfied when they can actively influence their search, although this may come along with higher interaction effort and cognitive load [34]. Besides, it has been shown that integrating RS with more interactivity improves, among others, transparency and perceived recommendation quality, which is more decisive than objective accuracy [66, 35, 51, 32]. Increased interactivity may be realized by using other preference elicitation methods than ratings and by eliciting preferences in an ad-hoc fashion, allowing users to immediately observe how their changes affect the results [18, 64, 42, 12]. A greater extent of control seems also beneficial for exploring large item spaces, especially when the search goal is vague [42], and for adapting recommendations towards situational needs. Further, interactive RS may help to alleviate the cold-start problem, and to support users in circumstances where they do not want a persistent profile to be applied, e.g. due to privacy concerns or because it belongs to a different person [64, 7, 42, 12].

Early examples for interactive RS are dialog-based and critique-based approaches. The latter allow users to criticize recommendations based on predefined item metadata [9]. This avoids the problem that users have to formulate their search goal up-front as it is necessary in dialog-based systems. Developments such as *MovieTuner* [64] build on this principle, but rely solely on user-generated content, in particular tags that can be weighted by the user to change the current result set. Other examples of interactive RS comprise *SmallWorlds* [18], *TasteWeights* [7], *SetFusion* [49] or *MyMovieMixer* [41]. These approaches to provide users with more control over the recommendations use manipulable graphs for influencing the underlying CF algorithms [18], interfaces for weighting the different datasources and algorithms in hybrid settings [7, 49], or faceted filtering blended with automated recommendation methods [41]. They all have shown to improve user engagement and overall satisfaction.

To allow users controlling the recommendation process at a more coarse-grained level than providing ratings for single items, these interactive approaches use, for example, tags [64, 12], automatically selected content attributes [7] or predefined item facets [41]. Preference elicitation thus becomes detached from actual items, which indeed has several advantages, but may also result in difficulties. It requires availability of adequate background data and highly depends on the possibility to categorize items among certain dimensions that actually matter to users. Also, mentally establishing a search goal so that preferences can be expressed with respect to specific item features may be non-trivial for users with little domain knowledge or in the beginning of a search task [25]. Only few approaches such as the one proposed in [65] allow users to define areas of interest directly inside the item space. This, however, seems to be a promising and natural way of expressing preferences without having to articulate them explicitly, and without the need to know and rate particular items.

Although several approaches use some kind of visualization, primarily to disclose the reasons for items to be suggested (e.g. [7, 49, 41]), they are typically independent of the more complex, especially map-based visualizations mentioned before. It thus seems promising to visualize the item space of CF recommenders together with user preferences in an integrated fashion by means of a map. This also opens the possibility of increasing user control, and, in particular, of letting users interactively specify their current interests with respect to entire item regions. In sum, an interactive landscape based on item space and user preferences has the potential of facilitating the establishment and manipulation of user profiles.

3D ITEM SPACE VISUALIZATION TO PRESENT AND MANIPULATE USER PREFERENCES IN CF

We propose a method that visualizes the item space together with user preferences as estimated by a model-based CF algorithm, as well as resulting recommendations. The underlying preference profile used to generate recommendations can be interactively set up by the user in cold-start situations and further manipulated in case such a profile already exists. With respect to the research questions posed at the beginning of this paper, the process which is also described in Figure 1 can be divided into the following main steps:

1. Visualize the entire item space as a 2D map and automatically select item samples to be displayed as representatives for the different regions.
2. Present the current user's preferences so that hills indicate areas of high interest, valleys areas of low interest, resulting in a 3D landscape.
3. Allow users to interactively change the elevation profile, this way manipulating the underlying model used to generate landscape as well as recommendations.

In the following, we will describe these steps in more detail.

RQ1: Visualizing and Sampling the Item Space

In order to visualize the item space, we solely rely on common user feedback as is typically used as background data in CF. By using rating data provided by all users, this step is independent of data availability for the current user. Nevertheless, one issue arising when using ratings to plot such a representation is data sparsity, since users typically rate only a small number of items out of the entire item set. Hence, it may be difficult to adequately calculate similarities between items, which is a prerequisite for many algorithms that map high-dimensional data onto low-dimensional spaces. However, it should be noted that although we use explicit ratings, our approach could in principle also be applied to implicit data, which is usually more dense. In either case, handling the large amount of data could lead to decreased efficiency of mapping algorithms. In addition, semantics inherent in these data may only hardly be exploited, potentially tempering quality of the item positioning, thus hindering users to understand the resulting map.

For these reasons, we introduce an intermediate step before plotting items on a 2D surface. In fact, we use a more abstract representation of items by exploiting their description through latent factors as derived by a standard MF algorithm (Figure 1, 1a), which has already been shown to be successful for "putting recommendations on a map" [17].

When using MF, the user-item-matrix $\mathbf{R} \in \mathbb{R}^{|U| \times |I|}$ that contains the raw rating data for all users $u \in U$ and items $i \in I$, is decomposed into two low-rank matrices, namely $\mathbf{P} \in \mathbb{R}^{|U| \times |f|}$ and $\mathbf{Q} \in \mathbb{R}^{|I| \times |f|}$, where f represents a predefined number of factors[2]. These matrices approximate the original user-item-matrix such that calculating the inner product of a user's factor vector $\vec{p}_u$ of $\mathbf{P}$ and an item factor vector $\vec{q}_i$ of $\mathbf{Q}$ returns the predicted rating $\hat{r}_{ui}$ for user u and item i. Estimating a user's ratings for all items is consequently done as follows:

$$\hat{r}_u = \vec{p}_u \mathbf{Q}^{\mathrm{T}} \tag{1}$$

By relying on a latent factor model, we take advantage of the fact that the factors implicitly convey semantics without requiring explicitly defined content data [37, 53, 13], Therefore, a mapping of the item space can be produced that is likely to be understood by users. Moreover, we circumvent any issues that may arise from sparsity, since MF can handle such matrices very efficiently [37]. Finally, by using a MF algorithm at this stage, we can draw on the derived user factor vector (Figure 1, 1a) also in the next step of the process to generate recommendations for the current user. Thereby, we take advantage of the fact that this widely used method is known for high recommendation quality [37, 36].

Next, we map the still high-dimensional item data onto a low-dimensional Euclidean space by using MDS [6, 28]. In order to visualize such data, different methods have been proposed [29, 28]. Typically, they rely on content information, so that the decision for a certain method depends on the item features. Geometric projections and scatter plots have been used very often for this purpose [56]. But, they can also be usefully applied when the dimensions are constructed by automated dimensionality reduction [10]. This is usually the case for datasets used by RS [2]. Thus, although other methods might be used, we chose MDS to calculate two-dimensional coordinates for all items. Using these coordinates, the resulting map visualization positions items based on their similarities (Figure 1, 1b). MDS ensures distances between any two items to be small if they are similar to each other, and large otherwise. We calculate the similarities used as input for the MDS algorithm by means of the Euclidean distance between item factor vectors $\vec{q}_i$, which seems reasonable since it naturally fits the positioning approach of MDS. As shown with the maps generated in [17], we assume that by relying on latent factor representations, it will adequately be reflected how items actually relate to each other. Thus, users should be able to perceive items close to each other as actually similar.

[2] Note that by setting $f = 2$ the item-factor-matrix could indeed be directly represented as a map. However, this would later result in reduced recommendation quality [37].

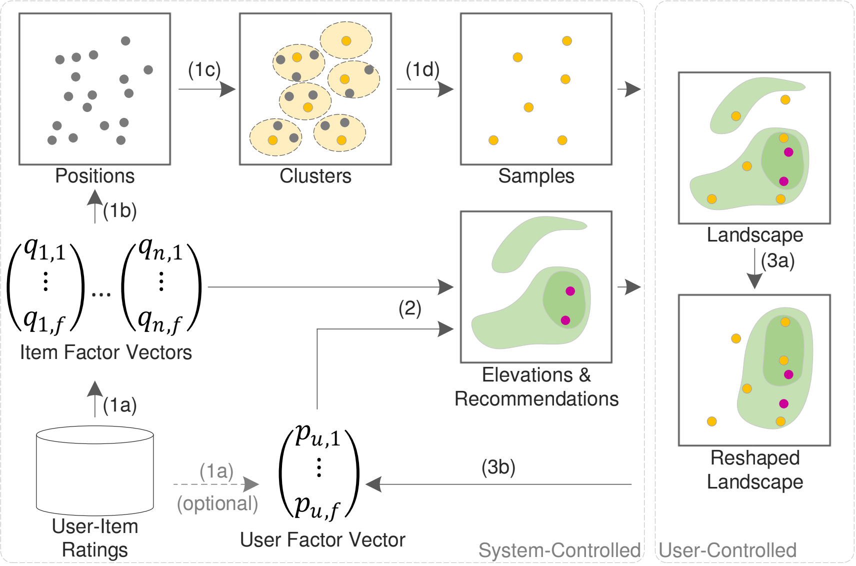


Figure 1. To generate the 3D item space visualization for presenting and manipulating preferences in CF, we start by using Matrix Factorization to obtain latent factors for users and items (1a). Item factor vectors then serve to determine the item positions (gray) using Multidimensional Scaling (1b). By applying k-means clustering (1c), popular representative items are chosen as samples (yellow) to be displayed on the resulting map (1d). In addition, item factor vectors are used to calculate predictions for the current user by taking his or her user factor vector into account (2). This results in elevations representing the user's preferences as well as in the actual recommendations (magenta). Note that the user factor vector is optional, so that this step is left out at cold-start where the user is instead presented with a flat surface. In any case, the user is then able to influence the recommendation process by reshaping the landscape, i.e. creating hills and valleys (3a), which finally leads to recalculation of the user factor vector (3b).

Now, in principle, items could already be plotted onto a 2D surface. However, due to the sheer mass of items, this is not a practical solution and would overwhelm users instead of providing an intuitive overview. Instead of showing popular items and additionally labeling certain areas of the map as in [17], we aim at generating an understanding for the item space based only on items themselves. Therefore, we select items that are representative for different regions of the quadratic map, and present them to the user. To perform this sampling, we use a *k-means* clustering [20] since we are in a Euclidean space (Figure 1, 1c). This allows us to control the trade-off between representativeness and number of items. Then, to determine a representative item for each cluster, we consider the five items that are closest to a cluster center and finally choose the most popular one as a sample, i.e. the item with highest number of ratings (Figure 1, 1d). Chosen items serve as representatives of the respective clusters and are at the same time likely to be known to many users. Based on the map dimensions and these sample items, we render an initial map.

RQ2: Presenting Preferences and Recommendations

In addition to the item space samples in the initial map, we also present the user's preferences and show recommended items in the context of the overall item space. Although the only items shown initially are the samples representing the different regions of the item space, in fact, all items have been assigned a certain position on the map. This allows us to exploit the third dimension by showing a landscape where the elevation indicates the system's predicted preferences among all items in respect to the current user. We therefore use the ratings predicted as usual by MF: Areas containing items with high predicted ratings are visualized as hills, those with low ratings as valleys. Since recommendations lie in areas where the system has predicted items to be of high interest, i.e. on hills, we assume the user will thus better understand how the system models his or her interests, and how this relates to actually expressed preferences as well as recommended items.

If the current user's preferences were previously elicited, e.g. by ratings, a latent factor vector for that user is already available. This vector $\vec{p}_u$ derived in the MF offline learning phase in the previous step (Figure 1, 1a) may now be used to calculate predictions as shown in (1) online. The resulting predictions $\hat{r}_u$ are used directly to select top-n items with highest scores as recommendations, but, as outlined above, also for setting up the elevation profile, and thus the 3D landscape. For this purpose, we linearly map the prediction for every item onto a height value, and consequently set the surface elevation at the item's respective position to this value. Then, to present the user with a visualization that actually resembles a landscape, the elevation of spaces between items is set to a level similar to adjacent items (otherwise, only spikes would appear at every item position). Therefore, we transfer height values of the items to their surrounding area where no items exist in a step-wise manner, decreasing with each step. Afterwards, we apply a Gaussian smoothing function and finally re-adjust the elevation at the actual item positions.

In case rating data for the current user are unavailable or the user does not want to apply an existing profile, i.e. we cannot use the user factor vector, the elevation profile is set to a neutral level. The visualization then shows a flat map surface and samples, both generated independently of the current user.

RQ3: Interactively Manipulating a Preference Profile

Regardless of whether the landscape already represents user preferences or just shows a flat surface when no user profile is available, the user can now interactively influence the underlying model, and consequently the recommendations. By shaping the landscape, i.e. raising or lowering the surface, the user is able to interactively express preferences for entire regions of the item space (Figure 1, 3a). The subsequent recalculation of predictions happens online—either continuously with each interaction, or when explicitly requested by the user, thus avoiding constant, possibly confusing updates of the visualization. In any case, we interpret changes to the landscape as user adjustments of the estimated ratings for the items that have led to the elevation of the respective areas. Note that this is independent of which items are actually shown, but instead takes all items in an area into account. The elevation values changed by the user are used to replace the rating predictions $\hat{r}_u$ calculated previously by new preference values, resulting in the vector $\vec{x}_u$. Based on this, we now set up a new user factor vector or recalculate the existing one by reformulating (1):

$$\vec{p}_u = \mathbf{Q}^+ \vec{x}_u, \tag{2}$$

where we use the pseudoinverse $\mathbf{Q}^+$ to approximate a solution via *Singular Value Decomposition* (since $\mathbf{Q}$ is non-quadratic). The updated user factor vector $\vec{p}_u$ is then fed back into the recommendations process (Figure 1, 3b), where it can be used to again predict ratings, leading to new recommendations as well as a new, adapted elevation profile. Thus, the user can immediately observe how the actions performed affect the recommender's results and the underlying preference profile.

DEMONSTRATOR

In this section, we present a demonstrator for our interactive recommending approach based on the 3D item space visualization described above. We implemented the demonstrator as a web-based application positioned within the movie domain. In addition to demonstration purposes, we also aimed at conducting user experiments with this demonstration system. In the following, we expand on its interaction concept and explain the implementation in more detail.

Interaction Concept

The user interface (Figure 2) is basically divided into four main parts: Working area (A), an area showing recommended items (B), detail area with information on the currently selected movie (C), and a palette of available interaction tools (D).

Within the working area, the visualization generated according to the steps described in the previous section is shown. In addition to the quadratic map surface representing the item space, the sample items, and the hills and valleys indicating the user's preferences, we color the surface to resemble a topographical map. Therefore, we use a function that assigns colors to particular levels of elevation while ensuring smooth transitions between them. This way, we aim at further facilitating the user's perception of the landscape and how it reflects the varying interests. Items are depicted with the help of movie posters directly on the map. Recommended items are additionally highlighted by means of a magenta-colored margin. Recommendations are also shown in the area at the bottom of the screen in form of a more conventional list. When the user hovers over an item on the map or in the recommendation list, the detail area is immediately updated and reveals further information on the respective movie (e.g. title, director, plot description and tags). Note that this content-related data is only used to provide users with additional information, and is not involved in the process of creating the visualization or generating recommendations. Finally, there is a palette showing several tools that may be used to perform interactions within the working area. Each tool has two functionalities that correspond to the left and right mouse button, respectively:

1. *Raise/Dig:* This tool can be used to shape the landscape, i.e. to create hills (left-click) and valleys (right-click) within the quadratic boundaries of the map. If selected, the mouse cursor shows a shovel icon and a small round white area surrounding the cursor indicates where the surface will be altered when clicking[3] (see also Figure 3). The highest or lowest possible elevation is thereby restricted through the linear mapping of predictions onto height values.
2. *Rotate/Pan:* As known from many 3D applications, this tool allows the user to rotate the entire perspective or to pan through the landscape.
3. *Show/Hide:* Inspired by [59], this tool helps to explore the item space in more detail. In case the user wants to see more than the initially shown samples, he or she can bring up additional items by left-clicking on the map (see also Figure 4). Then, the most popular of the five items closest to the cursor gets added. Right-clicking on an item already shown in turn removes this item from the map, which is particularly useful in case the map gets too crowded.

Independent of the tool currently selected, the user can always zoom in and out by using the mouse wheel.

Implementation Details

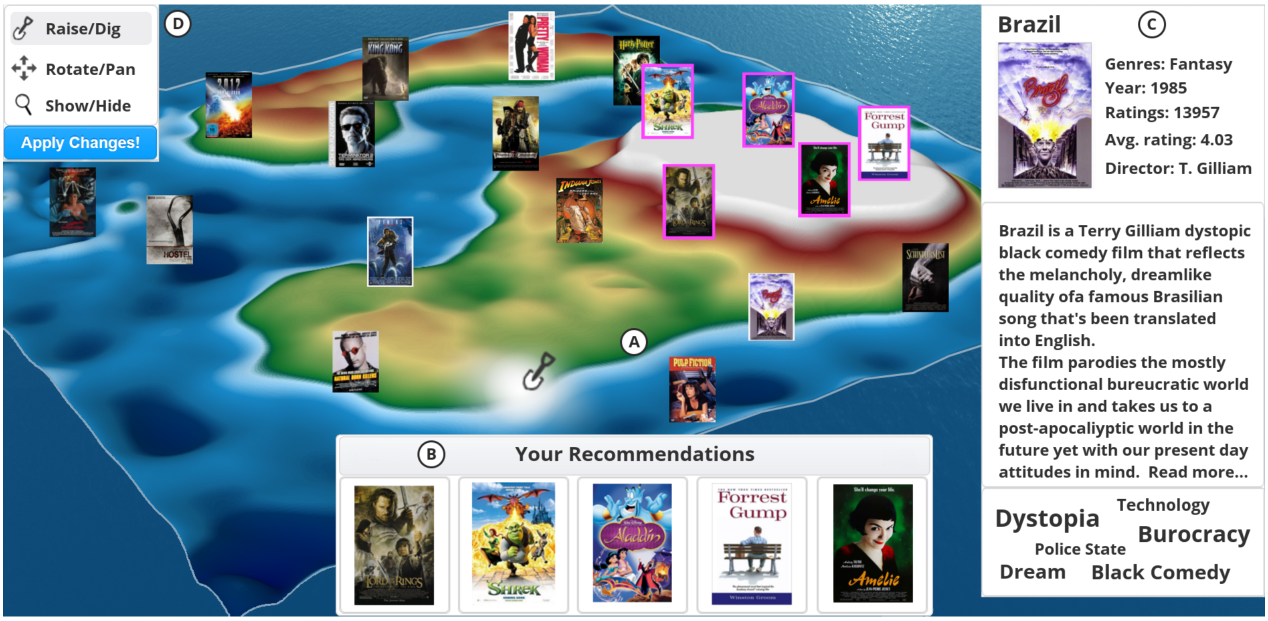
For implementing the process described, we first use the *Stochastic Gradient Descent* algorithm[4] from the *Apache Mahout*[5] library. This well-proven implementation of a standard MF algorithm allows us to derive the latent factor model used for calculating item similarities and rating predictions with performance up to standard (*RMSE* of 0.80 using 10-fold cross validation). As background data, we utilize the *MovieLens 20M Dataset*[6] containing about 20 million ratings from 137 000 users given to 27 000 movies. In principle, our approach may also be applied to other domains such as books, music, or any other type of commercial goods, in particular because CF, which is the underlying basis for our approach, is generally regarded as domain-independent. However, the MovieLens datasets are well-established within RS research, and, from our point of view, an appropriate means to show that our approach works as expected for *experience products*.

[3] Depending on the demonstrator's configuration, changes to the landscape are fed back into the recommendation process either continuously triggered by every mouse click, or only as soon as the user feels confident with the manipulations and uses the "Apply Changes"-button right underneath the palette (Figure 2, D). More details on how this is done can be found in the previous section.

[4] *ParallelSGDFactorizer* (8 factors, 16 iterations, $\lambda = 0.001$).

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

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



Brazil

Genres: Fantasy
Year: 1985
Ratings: 13957
Avg. rating: 4.03
Director: T. Gilliam

Brazil is a Terry Gilliam dystopic black comedy film that reflects the melancholy, dreamlike quality ofa famous Brasilian song that's been translated into English.
The film parodies the mostly disfunctional bureucratic world we live in and takes us to a post-apocalyptic world in the future yet with our present day attitudes in mind. Read more...

Figure 2. Screenshot of our demonstrator: Working area (A) visualizing the item space as a quadratic map that includes movie posters depicting the automatically chosen sample items and represents the user's preferences by the surface elevation; recommended items (B), which are also shown inside the landscape as posters highlighted by a magenta-colored margin; detail information on the currently selected movie (C); and a palette (D) of available interaction tools (in this example, the *Raise/Dig*-tool is selected, which appears at the position of the cursor in the lower middle part of the screen).

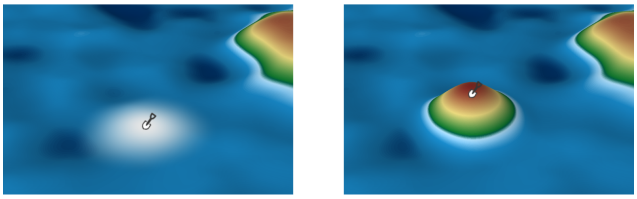


Figure 3. Using the *Raise/Dig*-tool, the user is able to shape the landscape, here by expressing his or her interest through forming a hill.

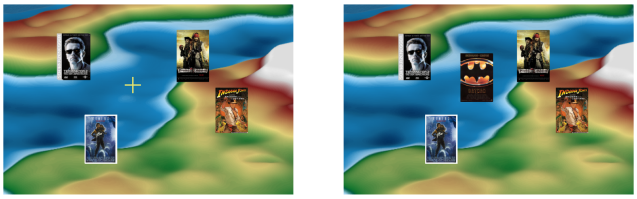


Figure 4. With the *Show/Hide*-tool, further items can requested to be shown in addition to the initially presented samples (inspired by [59]).

This way, we aim at ensuring a sufficient degree of ecological validity. Although not necessary for our 3D item space visualization in general, we additionally enrich the dataset by importing content-related information as well as movie posters from the *TMDb* website[7] in order to provide users with an appealing and informative presentation of the actual items.

Next, based on item factor vectors derived in the MF learning phase, we calculate item similarities which go into a MDS algorithm, resulting in the mapping used to arrange the items on the surface. Therefore, we rely on an implementation by the *Algorithmics Group*[8]. For clustering the items in order to determine representative samples, we use a *k-means* algorithm we implemented ourselves with $k = 30$. Early qualitative experiments suggested this number of initial samples to sufficiently represent the item space while not overwhelming users visually (Figure 2 is print-optimized and shows less samples). Finally, for visualizing the 3D landscape in our web-based application, we use the Javascript 3D library *three.js*[9].

EMPIRICAL USER STUDY

To evaluate our approach against the research questions, we conducted an empirical user study. We were particularly interested in examining the item space representation and the item sampling, the presentation of user preferences in form of a landscape, the interactive tools for shaping the surface, as well as the effect of these methods at cold-start and with an existing user profile. To assess the effectiveness of our approach, we constructed tasks that focus on these different aspects. We measured the user's perception of different system quality factors, especially with respect to subjective recommendation quality and perceived transparency as well as overall satisfaction and user experience.

Method

Participants and materials: We evaluated our approach using the demonstration system described in the previous section[10].

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

[8] http://algo.uni-konstanz.de/software/mdsj/

[9] https://threejs.org/

[10] The version of our demonstrator used in the study was slightly different than the one presented in this paper: The interface elements as well as the coloring of the landscape were more simple, and we used an earlier edition of the MovieLens dataset (the 10M version, see http://grouplens.org/datasets/movielens/10m/).

We recruited 32 (10 female) participants with age ranging from 18 to 34 ($M = 24.22$, $SD = 3.61$). The majority had a high school (62.5 %) or a university degree (34.4 %). Participants were asked to use the demonstrator under controlled conditions in a lab-based setting. They used a desktop PC with 24″ LCD (1920 × 1200 px resolution) and a common web browser to interact with the system and to fill in a questionnaire. The interface was in English, but participants (all non-native English-speakers) were explicitly allowed to ask the moderator for translations.

Tasks: The study was structured into three tasks, which were presented to each participant in the same order:

1. *Introductory search and exploration:* The first task can be seen as an introductory task focused on general exploration, orientation in the item space, and familiarizing with the interaction possibilities. In consecutive subtasks, participants were asked to explore the map in order to find three movies fulfilling following criteria: 1a) popular movies (more than 30 000 ratings), 1b) movies suitable for children, and 1c) movies directed by Quentin Tarantino. For each subtask, three minutes were given. Interaction was restricted to exploration, i.e. manipulating the landscape was not possible. No elevations and recommendations were present.
2. *Establishing a profile at cold-start:* In order to evaluate our system in cold-start situations, in this task, participants were asked to express their preferences on a flat surface, i.e. no profile was initially visualized. Starting from the flat surface, participants had to use the available tools to shape the landscape. Participants finished interaction at their own discretion, whereupon a new profile, and thus a user factor vector, was created. Resulting recommendations and the new landscape were presented afterwards to the user.
3. *Manipulating an existing profile:* This task addressed the situation where a user wants to manipulate an existing profile according to his or her current preferences. At the beginning of this task, the elevation profile was set up according to the preferences of an existing user[11]. Participants then had to alter the resulting landscape towards their own preferences. Recommendations and landscape, i.e. the elevations on the map, were updated continuously.

Questionnaires and log data: In order to assess the participants' subjective perception, we used a questionnaire that was primarily composed of different existing constructs[12]. At the beginning of each session, we elicited demographics and domain knowledge (regarding movies and 3D applications). Then, subsequent to task 2, we assessed perceived recommendation quality [33], transparency [50], interaction effort [33] and interaction adequacy [50]. We complemented these existing constructs with a few questionnaire items generated by ourselves, primarily regarding aspects very specific to our approach (e.g. comprehensibility of the landscape and the positions of recommended items inside, perceived controllability of the recommendation process). Next, following task 3, we used the same constructs again, but also assessed perceived control [50] and used self-generated items concerning manipulation of existing profiles. Finally, at the end of each session, we asked participants some questions regarding their general impression of the system. For this, we used constructs such as system effectiveness [33] and perceived usefulness [50], as well as some additional self-generated items. Across tasks, this resulted in about 75 items. In addition, to measure usability, user experience and engagement with the system, we applied the *System Usability Scale* (SUS) [8], *User Experience Questionnaire* (UEQ) [39] and subscales of *Intrinsic Motivation Inventory* (IMI) [54]. All items were assessed on a positive 5-point Likert scale, except the ones from UEQ (7-point bipolar scale) and IMI (positive 7-point Likert scale)[13].

In each session, we also logged interaction behavior, i.e. actions such as selecting tools, shaping the landscape (and how long this took), or showing/hiding items. In addition, we measured task times and, especially for task 1, recorded whether participants were able to accomplish the respective task.

Results

Overall, participants were very satisfied with the system ($M = 3.94$, $SD = 0.76$) and enjoyed using it ($M^{\dagger} = 5.46$, $SD = 0.97$). They perceived the recommender as effective ($M = 3.72$, $SD = 0.74$) and useful ($M = 3.75$, $SD = 0.76$). Table 2 (General results) shows the results for some selected questionnaire items from these general constructs, that particularly emphasize the overall quality of recommendations and the user's enjoyment when using the demonstrator.

Table 1 illustrates results with respect to the general constructs perceived recommendation quality, transparency, interaction effort and interaction adequacy, which we assessed after task 2 and 3, respectively.

	Task 2		Task 3		
	M	*SD*	*M*	*SD*	*d*
Perceived rec. quality	3.57	0.89	3.89	0.60	.42
Transparency	3.91	1.09	3.63	1.07	.26
Interaction effort*	3.75	0.76	3.21	0.93	.64
Interaction adequacy	3.47	0.88	3.61	0.90	.16

Table 1. Differences between task 2 and 3 with respect to perception of recommendations and the interaction (* marks the only construct yielding a significant difference, *d* represents Cohen's effect size value).

In the following, we address our three research questions by expanding on these general constructs and, in particular, by presenting further specific results.

RQ1: Visualizing and sampling the item space: Participants predominantly agreed with the statement that the item positioning on the map was comprehensible and found the landscape helpful for obtaining an overview of the entire item space. Consequently, this facilitated their awareness of possible choice options. Table 2 (RQ1) summarizes the descriptive statistics.

When participants were asked to explore the item space in order to find movies fulfilling different criteria in the subtasks

[11] We carefully selected three existing user profiles from the underlying MovieLens dataset, all very different to each other. Out of these profiles, one was randomly chosen for each participant. Then, in the following, we used the corresponding user factor vector.

[12] We translated questionnaire items to present them in German language, sometimes with slightly adapted formulations.

[13] Mean values of such items are in the following indicated as $M^{\dagger}$.

of task 1, all of them were able to find three movies suitable for children within the given time limit (1b). Popular movies could still be successfully found by 88 % (1a), while only 56 % found three movies directed by Quentin Tarantino (1c). This is also reflected in the time participants needed to accomplish the subtasks: Using a one-factorial RM-ANOVA, we found significant differences, $F(2,62) = 32.801$, $p = .000$. Post hoc comparisons using Bonferroni correction show no difference between task 1a ($M = 1.46$ min, $SD = 0.88$) and task 1b ($M = 1.25$ min, $SD = 0.60$). However, the subtask of finding movies directed by Tarantino took significantly more time ($M = 2.70$ min, $SD = 0.88$) than the two others ($p < .01$).

When we asked participants whether they would use the system in different search situations, we again found significant differences, $F(1.38,43.01) = 21.010$, $p = .000$. Participants rated the system as very useful for situations where they would have *no* ($M = 3.91$, $SD = 1.25$) or only a *vague* search goal in mind ($M = 4.06$, $SD = 0.98$). Here, post hoc comparisons denote no significance difference. In contrast, participants stated that they would use the system significantly less likely ($p < .01$) in situations with a *concrete* search goal, i.e. for known-item search ($M = 2.44$, $SD = 1.48$).

RQ2: Presenting preferences and recommendations: As already presented in Table 1, perceived recommendation quality and transparency were assessed very positively (no significant differences between task 2 and 3). In addition, the items more specific to our approach reported in Table 2 (RQ2) confirm that the landscape helped participants to understand how their preference profile was represented within the system and that they understood why items had been recommended.

To compare the two tasks, we also assessed the comprehensibility of the generated landscape, i.e. the elevations on the map representing the estimated preferences. We found a significant difference ($t(31) = 2.37$, $p < .05$). In task 2, were participants started from a flat surface, they stated that they understood why the landscape was finally generated the way it was ($M = 3.94$, $SD = 0.91$). When manipulating a profile from another person in task 3, the comprehensibility was rated lower ($M = 3.41$, $SD = 1.04$). Cohen's d, however, suggests only a moderate effect size ($d = .54$).

RQ3: Interactively manipulating a preference profile: The general construct assessing perceived control over the system yielded satisfying results ($M = 3.54$, $SD = 0.94$). When looking at specific questionnaire items regarding the quality of the interaction possibilities provided to express preferences (Table 2, RQ3), scores were even better: Participants felt to be able to tell the system what they like/dislike, i.e. in our case to create hills and valleys, in cold-start situations (assessed after task 2), and to modify an existing preference profile (assessed after task 3). Overall, participants felt in control over the recommendation process by manipulating the landscape.

With respect to perceived interaction effort, Table 1 shows the overall positive results for our system. However, we found a significant difference between task 2 and 3 ($t(31) = 3.76$, $p < .01$) with medium effect size. This was not reflected in the time participants needed to accomplish the tasks: Both took a statistically similar amount of time, $M = 6.48$ min ($SD = 2.40$) for task 2, and $M = 5.53$ min ($SD = 3.48$) for task 3, with a rather small effect size ($d = .32$).

In general, as shown in Table 1, interaction adequacy was assessed equally positive for both tasks (small effect size). When asked specifically whether they understood how their interactions affected the landscape, participants seemed also satisfied, in task 2 ($M = 4.00$, $SD = 0.76$) and in task 3 ($M = 3.44$, $SD = 1.24$), without significant difference ($d = .54$).

Also the results for the constructs mentioned before, e.g. overall satisfaction, system effectiveness, perceived usefulness and recommendation quality, show a positive assessment of the interaction possibilities for manipulating a preference profile.

Usability and user experience: Usability of our demonstrator was evaluated as *good* with a SUS-score of 75. On the different scales of the UEQ, we received promising results (ranging from 0.84 to 2.10), in particular, for perspicuity (1.66, *good*), stimulation (1.56, *excellent*) and novelty (2.10, *excellent*).

Demographics and domain knowledge: Participants generally stated that they love movies ($M = 3.91$, $SD = 0.78$), and 75 % reported that they regularly use sites like *IMDb* or *Rotten Tomatoes* for searching further information. Participants were not very familiar with standard 3D applications ($M = 2.25$, $SD = 0.62$), e.g. *Google Earth* or 3D computer games. The expertise with professional 3D applications such as *3DS Max* was, as expected, even lower ($M = 1.41$, $SD = 0.61$). However, we did not find any noteworthy influence of demographics or domain knowledge on our dependent variables.

Discussion

Overall, the study results suggest that our approach provides users with an easy to understand 3D visualization of item space and preferences. Although we built on model-based CF, which is generally considered to be a rather opaque technique, participants were able to make sense of the generated map and the positioning of items on the map. Relying only on the hidden semantics of latent factors, the initial selection of representative samples appeared to be a good starting point for further exploration. Observed interaction behavior shows that the interaction tools provided, e.g. the possibility to request more items, also contribute to participants quickly getting an overview. Consequently, they were able to successfully accomplish search tasks although the overall number of items in the dataset was large. As expected, our approach performed better in situations with a more general search direction in mind than for known-item search. Looking for concrete items could however easily be supported by providing additional search functionalities. Our study, in contrast, has shown that using a latent factor model without any content information seems especially of value in different, yet very common situations where users are searching with respect to "soft" criteria.

With respect to representation of their preferences, as well as the means provided to establish or manipulate the underlying profile, participants were also satisfied. Using the elevation profile of the map to visualize the user's preferences seemed to be supportive in order to reveal how the user is represented within the system. Furthermore, this way being able to set

	Item	*M*	*SD*
RQ1	The positioning of movies inside the landscape was comprehensible.	3.31	0.90
	The presentation of movies inside the landscape helped me getting an overview of the item space.	3.91	0.93
	The recommender system makes me more aware of my choice options.	3.91	0.93
RQ2	The landscape helped me to understand my user profile within the system.	3.63	1.07
	I think the landscape helped me to understand why the movies have been recommended to me.	3.69	0.93
RQ3	The recommender system allows me to tell what I like/dislike.	3.97	1.12
	The recommender system allows me to modify my taste profile.	3.72	0.96
	I felt to be in control over the recommendation process by manipulating the landscape.	3.69	0.97
General results	The recommender system gave me valuable recommendations.	3.97	0.90
	The recommender system helped me find the ideal item.	4.00	0.76
	I enjoyed using the system very much.†	5.47	1.02
	Using the system was fun to do.†	5.38	1.26

Table 2. Mean values and standard deviations for selected questionnaire items, grouped by our research questions († indicates items assessed using a positive 7-point Likert scale, in all other cases, a positive 5-point Likert scale was used).

recommended items in relation to the rest of the item space, appears to positively influence perceived recommendation transparency. This is reflected by the fact that participants stated to understand why recommendations were shown at certain positions. The moderate significant difference regarding comprehensibility of the landscape between task 2 and 3 is likely a result of initially presenting a profile from another person in the latter case. Since preferences from a completely different profile might still have influenced the results after participants finished the interaction, it seems reasonable that they perceived the landscape as slightly less comprehensible.

Explicit user feedback is often very sparse and motivating users to state their preferences, for example, by means of ratings, is known to be difficult [19, 16]. In this light, it seems particularly promising that participants felt in control over our system while enjoying the interaction. Thus, shaping the landscape by creating hills and valleys appears to be an appropriate means for expressing preferences. This is also supported by the positive results in terms of interaction adequacy, usability, and user experience. The significant difference between task 2 and 3 with respect to perceived effort may again be ascribed to the fact that participants had to manipulate an existing profile in the latter case, which is likely to be a more complex task than starting from a flat surface. In addition, the continuous updates of landscape and recommendations in task 3 may also have contributed to perceiving the effort to be slightly higher. Thus, although the scores are still in a satisfactory range for both tasks, further investigation will be needed to account for the different task settings. Either way, participants were satisfied with the resulting recommendations, both when they started to establish a preference profile as well as when they had to manipulate an existing one.

CONCLUSIONS AND OUTLOOK

To answer the research questions posed at the beginning of this paper, we introduced a novel 3D visualization with a landscape of hills and valleys in order to represent a large item space, show the user's preferences in that space, and allow him or her to manipulate the underlying model. We implemented a demonstrator that indicates the usefulness of our approach—also in cold-start situations. In the user study we conducted, we obtained promising results concerning our research questions, and especially regarding perceived transparency, recommendation quality, user enjoyment, and degree of control users are able to exert over the system.

Apparently, a latent factor model inferred by MF from ratings as they are customary in CF may not only serve to calculate accurate recommendations, but also conveys semantics that can be revealed to the user. While this is in line with earlier research [53, 13], we show that latent factors may be a legitimate source for positioning a large number of items on a map that users perceive as comprehensible. Without requiring any content-related data, preferences can both be presented and successfully elicited with respect to regions of the item space the user is particularly interested in—independent of knowing and rating specific items. Although MF-based methods are typically intransparent due to their statistical nature, our study suggests that using a modern visualization technique together with representative sample items, supports users in understanding the representation of their preferences within the system, i.e. the user model, and the resulting recommendations.

Despite the potential shown by our interactive recommending approach based on conventional model-based CF, it is in principle independent of algorithms and background data. In future work, we therefore aim at using recommender algorithms other than MF, mapping and sampling techniques besides MDS and k-means, and also further datasources, e.g. content information instead of or in addition to ratings. This goes along with our goal of implementing the approach in a different domain or in a cross-domain scenario, where the need to deal with more heterogeneous data as well as a larger number of items is even more apparent. Furthermore, there is room left for improvement with respect to the visualization and interaction concept. For instance, additional samples could immediately be shown when zooming in. Also, the usage of a map metaphor may be further exploited, e.g. by highlighting regions on the map and labeling them with tags. In general, one can think of using entirely different interaction mechanisms or even a tangible user interface. Finally, while the present user study focused on a proof-of-concept, we are also interested in conducting more in-depth comparisons, in particular with a baseline system as well as other state-of-the-art interactive RS and visualizations.

REFERENCES

1. Jae-Wook Ahn and Peter Brusilovsky. 2013. Adaptive Visualization for Exploratory Information Retrieval. *Information Processing & Management* 49, 5 (2013), 1139–1164.
2. Xavier Amatriain and Josep M. Pujol. 2015. *Recommender Systems Handbook*. Springer US, Chapter Data Mining Methods for Recommender Systems, 227–262.
3. Xavier Amatriain, Josep M. Pujol, Nava Tintarev, and Nuria Oliver. 2009. Rate It Again: Increasing Recommendation Accuracy by User Re-rating. In *Proc. RecSys '09*. ACM, 173–180.
4. Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. In *Proc. IUI '13*. ACM, 49–56.
5. Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. 2013. Semantic Audio Content-Based Music Recommendation and Visualization Based on User Preference Examples. *Information Processing & Management* 49, 1 (2013), 13–33.
6. Ingwer Borg and Patrick J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications* (2 ed.). Springer.
7. Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proc. RecSys '12*. ACM, 35–42.
8. John Brooke. 1996. SUS – A Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*. Taylor & Francis, 189–194.
9. Li Chen and Pearl Pu. 2012. Critiquing-Based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150.
10. Mei C. Chuah. 1998. Dynamic Aggregation with Circular Visual Designs. In *Proc. INFOVIS '98*. IEEE, 35–43.
11. Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2012. User Effort vs. Accuracy in Rating-Based Elicitation. In *Proc. RecSys '12*. ACM, 27–34.
12. Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2016a. Tag-Enhanced Collaborative Filtering for Increasing Transparency and Interactive Control. In *Proc. UMAP '16*. ACM, 169–173.
13. Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2016b. Towards Understanding Latent Factors and User Profiles by Enhancing Matrix Factorization with Tags. In *Poster Proc. RecSys '16*.
14. Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2014. Active Learning Strategies for Rating Elicitation in Collaborative Filtering: A System-Wide Perspective. *ACM Transactions on Intelligent Systems and Technology* 5, 1 (2014), 13:1–13:33.
15. Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion Space: A Scalable Tool for Browsing Online Comments. In *Proc. CHI '10*. ACM, 1175–1184.
16. Sebastian Feil, Martin Kretzer, Karl Werder, and Alexander Maedche. 2016. Using Gamification to Tackle the Cold-Start Problem in Recommender Systems. In *CSCW '16 Companion*. ACM, 253–256.
17. Emden Gansner, Yifan Hu, Stephen Kobourov, and Chris Volinsky. 2009. Putting Recommendations on the Map – Visualizing Clusters and Relations. In *Proc. RecSys '09*. ACM, 345–348.
18. Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. 2010. SmallWorlds: Visualizing Social Recommendations. *Computer Graphics Forum* 29, 3 (2010), 833–842.
19. F. Maxwell Harper, Xin Li, Yan Chen, and Joseph A. Konstan. 2005. An Economic Model of User Rating in an Online Recommender System. In *Proc. UM '05*. Springer, 307–316.
20. John A. Hartigan and M. Anthony Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.
21. Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive Recommender Systems: A Survey of the State of the Art and Future Research Challenges and Opportunities. *Expert Systems with Applications* 56, 1 (2016), 9–27.
22. Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.
23. Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. 2010. A Tour Through the Visualization Zoo. *ACM Queue* 53, 6 (2010), 59–67.
24. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proc. CSCW '00*. ACM, 241–250.
25. Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. *Recommender Systems Handbook*. Springer US, Chapter Human Decision Making and Recommender Systems, 611–648.
26. Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (2014), 8:1–8:26.
27. Yucheng Jin, Karsten Seipp, Erik Duval, and Katrien Verbert. 2016. Go With the Flow: Effects of Transparency and User Control on Targeted Advertising Using Flow Charts. In *Proc. AVI '16*. ACM, 68–75.

28. Martijn Kagie, Michiel van Wezel, and Patrick J. F. Groenen. 2011. *Recommender Systems Handbook*. Springer, Chapter Map Based Visualization of Product Catalogs, 547–576.

29. Daniel A. Keim and Hans-Peter Kriegel. 1996. Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on Knowledge and Data Engineering* 8, 6 (1996), 923–938.

30. Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North (Eds.). 2008. *Information Visualization: Human-Centered Issues and Perspectives*. Springer.

31. Mohammad Khoshneshin and W. Nick Street. 2010. Collaborative Filtering via Euclidean Embedding. In *Proc. RecSys '10*. ACM, 87–94.

32. Bart P. Knijnenburg and Martijn C. Willemsen. 2015. *Recommender Systems Handbook*. Springer US, Chapter Evaluating Recommender Systems with User Experiments, 309–352.

33. Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

34. Jürgen Koenemann and Nicholas J. Belkin. 1996. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proc. CHI '96*. ACM, 205–212.

35. Joseph A. Konstan and John Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123.

36. Yehuda Koren and Robert Bell. 2015. *Recommender Systems Handbook*. Springer US, Chapter Advances in Collaborative Filtering, 77–118.

37. Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.

38. Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 2015. 3D-Visualisierung zur Eingabe von Präferenzen in Empfehlungssystemen [3D visualization to elicit preferences in recommender systems]. In *Proc. M&C '15*. De Gruyter Oldenbourg, 123–132.

39. Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*. Springer, 63–76.

40. Benedikt Loepp, Catalin-Mihai Barbu, and Jürgen Ziegler. 2016. Interactive Recommending: Framework, State of Research and Future Challenges. In *Proc. EnCHIReS '16*.

41. Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *Proc. CHI '15*. ACM, 975–984.

42. Benedikt Loepp, Tim Hussein, and Jürgen Ziegler. 2014. Choice-Based Preference Elicitation for Collaborative Filtering Recommender Systems. In *Proc. CHI '14*. ACM, 3085–3094.

43. Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. 2003. Interfaces for Eliciting New User Preferences in Recommender Systems. In *Proc. UM '03*. Springer, 178–187.

44. Afshin Moin. 2014. A Unified Approach to Collaborative Data Visualization. In *Proc. SAC '14*. ACM, 280–286.

45. Chris Muelder, Thomas Provan, and Kwan-Liu Ma. 2010. Content Based Graph Visualization of Audio Data for Music Library Navigation. In *Proc. ISM '10*. IEEE, 129–136.

46. Sayooran Nagulendra and Julita Vassileva. 2014. Understanding and Controlling the Filter Bubble Through Interactive Visualization: A User Study. In *Proc. HT '14*. ACM, 107–115.

47. Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding From You*. Penguin Press.

48. Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: Delivering Multiple Aspects of News to Mitigate Media Bias. In *Proc. CHI '09*. ACM, 443–452.

49. Denis Parra, Peter Brusilovsky, and Christoph Trattner. 2014. See What You Want to See: Visual User-Driven Approach for Hybrid Recommendation. In *Proc. IUI '14*. ACM, 235–240.

50. Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proc. RecSys '11*. ACM, 157–164.

51. Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355.

52. Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2015. *Recommender Systems Handbook* (2 ed.). Springer US.

53. Marco Rossetti, Fabio Stella, and Markus Zanker. 2013. Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems. In *Proc. DEXA '13*. 162–167.

54. Richard M. Ryan. 1982. Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *Journal of Personality and Social Psychology* 43, 3 (1982), 450–461.

55. Amit Sharma and Dan Cosley. 2013. Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems. In *Proc. WWW '13*. ACM, 1133–1144.

56. Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. VL '96*. IEEE, 336–343.

57. Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts*. ACM, 830–831.

58. E. Isaac Sparling and Shilad Sen. 2011. Rating: How Difficult is it?. In *Proc. RecSys '11*. ACM, 149–156.

59. Pieter Jan Stappers, Gert Pasman, and Patrick J. F. Groenen. 2000. Exploring Databases for Taste or Inspiration with Interactive Multi-Dimensional Scaling. *Proc. HFES '00* (2000), 575–578.

60. Nava Tintarev and Judith Masthoff. 2015. *Recommender Systems Handbook*. Springer US, Chapter Explaining Recommendations: Design and Evaluation, 353–382.

61. James Uther and Judy Kay. 2003. VlUM, a Web-Based Visualisation of Large User Models. In *Proc. UM '03*. Springer, 198–202.

62. Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proc. IUI '13*. ACM, 351–362.

63. Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proc. IUI '09*. ACM, 47–56.

64. Jesse Vig, Shilad Sen, and John Riedl. 2011. Navigating the Tag Genome. In *Proc. IUI '11*. ACM, 93–102.

65. David Wong, Siamak Faridani, Ephrat Bitton, Björn Hartmann, and Ken Goldberg. 2011. The Diversity Donut: Enabling Participant Control over the Diversity of Recommended Responses. In *CHI '11 Extended Abstracts*. ACM, 1471–1476.

66. Bo Xiao and Izak Benbasat. 2007. E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly* 31, 1 (2007), 137–209.

The following article is reused from:

Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. Newsviz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 126–135. ACM, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394869

NewsViz: Depicting and Controlling Preference Profiles Using Interactive Treemaps in News Recommender Systems

Johannes Kunkel
University of Duisburg-Essen
Duisburg, Germany
johannes.kunkel@uni-due.de

Claudia Schwenger
University of Duisburg-Essen
Duisburg, Germany
claudia.schwenger@stud.uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

News articles are increasingly consumed digitally and recommender systems (RS) are widely used to personalize news feeds for their users. Thereby, particular concerns about possible biases arise. When RS filter news articles opaquely, they might "trap" their users in *filter bubbles.* Additionally, user preferences change frequently in the domain of news, which is challenging for automated RS. We argue that both issues can be mitigated by depicting an interactive version of the user's preference profile inside an overview of the entire domain of news articles. To this end, we introduce *NewsViz*, a RS that visualizes the domain space of online news as treemap, which can interactively be manipulated to personalize a feed of suggested news articles. In a user study ($N = 63$), we compared *NewsViz* to an interface based on sliders. While both prototypes yielded high results in terms of transparency, recommendation quality and user satisfaction, *NewsViz* outperformed its counterpart in the perceived degree of control. Structural equation modeling allows us to further uncover hitherto underestimated influences between quality aspects of RS. For instance, we found that the degree of overview of the item domain influenced the perceived quality of recommendations.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → Human computer interaction (HCI)**; **Treemaps**.

KEYWORDS

News Recommender Systems; Interactive Recommending; Information Visualization; Treemaps; Structural Equation Modeling

ACM Reference Format:
Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. 2020. NewsViz: Depicting and Controlling Preference Profiles Using Interactive Treemaps in News Recommender Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340631.3394869

1 INTRODUCTION

The domain of news is currently subject to a substantial change as many roles that were formerly undoubtedly associated with humans are now increasingly performed by machines. A prominent example for such roles is the curation of online news in which recommender systems (RS) progressively act as *gatekeepers* and decide which articles will be included in a personalized news feed and which will not [43, 44]. While RS in general have become quite accurate in computing personalized recommendations, it has been argued that accuracy is only one of many quality aspects of a RS [4, 35, 40].

When content is pro-actively personalized, users might overestimate how representative recommendations for the entire item domain are and become trapped in *filter bubbles* [45] or *echo chambers* [14]. Possible results include ideological segregation [14], burgeoning populism [10] and distribution of conspiracy theories [8]. While it remains under discussion whether algorithmic filtering is the main reason for filter bubbles [18, 41], incidents like the scandal around Cambridge Analytica have resulted in a broad public interest of algorithmic transparency and filter bubble effects [19]. One way to tackle such concerns is making users aware of not only the recommendations but also of items the algorithm omits [42].

Consequently, one task of RS can be defined as letting users *explore or understand the item space* [27]. Providing such a broad *overview* was observed to make users aware of blind-spots in their profile [55], help them to develop new preferences [39], and increase control over the recommendations [37]. A prominent way of conveying an overview of the item space is to present it as a scatterplot or geographical map [16, 37, 50].

While map-based visualizations have proven to be able to foster overview in RS, they are prone to visual clutter and seldom make efficient use of the entire available screen space [29]. More space-efficient in this regard are *treemaps* [51]. Treemaps are used to visualize tree structures as map in which tree nodes are represented as cells. Besides other domains, visualizations based on treemaps have also been applied to RS [31, 48]. Richthammer and Pernul, for instance, utilize an interactive treemap to cluster recommendations regarding their content, thus aiding users in comprehending their own situational needs [48]. However, the potential of treemaps could even be exploited more extensively, when not only recommendations are visualized but also their relation to the user's preferences in context of the entire item space. Aligning this with interactive methods for treemaps [1] could also raise the users' control over their recommendations.

In this paper, we combine the above: We introduce *NewsViz*, a novel news RS that utilizes a treemap with interactive cells to provide an overview of the item domain, visualize the user's preference profile, and let this profile interactively be adjusted. In this

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394869

way we aim at making the RS more transparent, reducing the risk of filter bubble effects, and enable control over recommendations. Summarized, we seek to answer the following research questions:

RQ1 How can a treemap visualization be leveraged to create an interactive control panel for RS?

RQ2 How are the aspects *overview*, *transparency*, and *control* influenced by the form of visualizing user preferences? How do they influence each other?

RQ3 What are possible benefits of treemap visualizations in terms of preventing filter bubble effects?

To answer these questions, we implemented *NewsViz* and conducted a user study that reveals high potential of treemaps to convey overview, transparency, and control in news RS. To investigate the effect of the treemap isolated from other aspects such as the recommendation algorithm, we compared *NewsViz* to a baseline system that replaces the treemap with an equally powerful interaction concept based on slider widgets. While both applications scored relatively high on the perceived degree of overview and transparency, we found that *NewsViz* outscored the slider-based prototype in terms of the degree of perceived control. Using structural equation modeling enables us to study the interdependencies of these constructs further, disclosing the particularly influential role of overview; not only on perceived transparency but also on less obvious aspects such as the system's effectiveness and the perceived quality of recommendations.

To summarize, we make the following contributions: (1) We demonstrate that treemaps can be used to control news RS more effectively compared to common slider widgets. (2) We underline the so far neglected importance of overview for RS and reveal implicit influences on aspects such as transparency and the perceived recommendation quality.

2 RELATED WORK

News articles are consumed increasingly digitally, thus lowering entry barriers for news distributors and costs of news dissemination. As the number of available news articles is growing, RS increasingly act as *gatekeepers* that automatically curate personalized news feeds [43, 44]. As a result, RS in the domain of news have a special responsibility since they can "exercise power over individuals" [9] in form of influencing the direction of their readers' awareness [9, 56]. But also from an algorithmic perspective, recommending news is particularly challenging: new items emerge constantly on arrival of recent news stories and user's preferences change frequently, sometimes even during the course of a day [30].

While providing control over recommendations is obviously particularly helpful in situations where user preferences vary often, surprisingly little research pursues to increase control in news RS. One exception is the qualitative study conducted by Harambam et al. [19], in which the authors observed that users of news RS strongly desire to have advanced means of interaction with the system. Harambam et al. mostly utilized sliders, which participants perceived as easy to use and "quite straightforward" but also as prone to overcrowd interfaces.

Outside the news domain, control of RS has been discussed in greater depth [21, 22]. Exerting control has been, beside others, related to supporting users to explore the item space [57, 58], influencing user satisfaction [12, 21, 28], and increasing trust into the RS [11]. Control in RS seems to be firmly tied to transparency of recommendations [22, 58], as users need to be educated of how to effectively influence recommendations.

Most often RS appear as *black boxes* to their users, as it remains opaque what data is used for personalization, why certain items are suggested, and how they relate to the user's preferences. Users who encounter such black box algorithms may react with distrust and may be reluctant to accept recommendations [25]. When transparency of a RS is low, it might also happen that users loose awareness of item space areas that are not recommended to them. Such users are effectively trapped in *filter bubbles* [45], which have been related to several negative societal consequences [8, 10, 14] and to a general threat for human creativity [33, 38]. However, others come to the conclusion that such concerns about filter bubbles are mostly exaggerated [18] or that filter bubbles are not primarily the result of RS but deliberately created by users themselves [2]. Nonetheless, algorithmic transparency and issues like filter bubbles have reached high visibility in the public and thus raised the demand for more transparent algorithms [19].

Opacity in RS can be tackled in various ways, for instance, by generating textual explanations for recommendations [25, 54] or by utilizing information visualization methods such as two-dimensional maps [16, 37] or Venn diagrams [46]. Accordingly designed transparent interfaces have shown potential to increase user satisfaction [17, 58] and may also help to educate users about their preferences thus facilitating self-reflection [33]. Following the same argumentative line, it was shown that when users feel educated about algorithmic workings of a RS, they can be more motivated to explore items beyond their usual interests [5], which helps them to receive more diverse recommendations [24] and have less blind-spots regarding the item space [20, 55]. In this sense, a broad, diverse, and unbiased viewpoint is a crucial necessity for democracy [23]. Not surprisingly, Jannach and Adomavicius list "help users to explore or understand the item space" as one of the purposes of a RS [27].

Such a comprehension of the item space can, for instance, be conveyed to users by utilizing visualizations based on maps or scatterplots [e.g. 16, 37, 50]. Corresponding visualizations have shown potential to increase transparency and user engagement [13, 15]. By leveraging the inherent comprehensibility of spatial relations, maps can help putting the user's preference model into context and thus letting recommendations become intuitively understandable [32]. A more abstract form of maps are *treemaps* [51], which are more space efficient as maps based on scatterplots [29]. Treemaps have also been applied to the domain of RS [31, 48].

Katarya et al. [31] and Richthammer and Pernul [48] utilize treemaps to depict the predicted fit of recommended movies in accordance to the active user's preferences. In particular, each treemap cell contains a recommendation and the size of those cells indicates how high the predicted rating of the corresponding item is. The prototype of Richthammer and Pernul adds a second hierarchical level, which groups recommendations regarding their movie genre. This visualization is accompanied with checkboxes that let users filter movies. Chang et al. [6] introduce a treemap-like interface for scrutinizing textual restaurant reviews. For different search queries, users can create separate treemaps. Each cell of

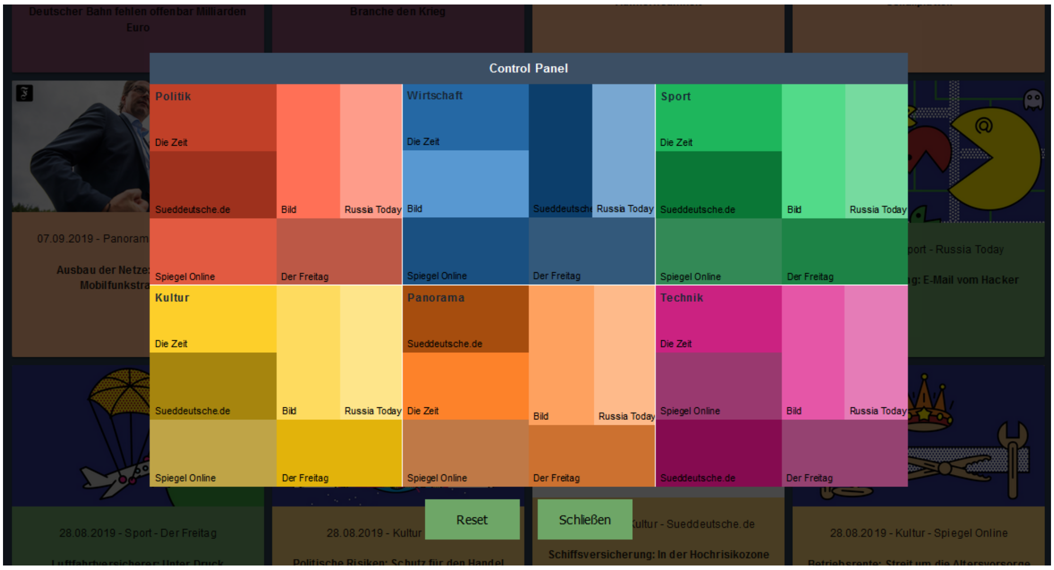


Figure 1: Screenshot of the *NewsViz* system. The user can hover any category or source inside the treemap visualization and scroll the mouse wheel up and down to enlarge or, respectively, shrink the corresponding cell. After finishing interaction, the user can click on a designated button or anywhere outside the panel to minimize it. In the background, the personalized news feed is updated with every interaction, thus providing immediate visual feedback to the user.

these treemaps corresponds to a keyword of the query. Relevance of keywords can be determined by users in changing the size of the corresponding treemap cell. Results of a user study show increased interaction quantity and general user satisfaction compared to a baseline system. Finally, treemaps have been used to display news in the commercial application *Newsmap*[1]. *Newsmap* demonstrates impressively how a large number of news can be displayed using a treemap.

3 *NEWSVIZ*: A TREEMAP TO CONTROL RECOMMENDATIONS OF ONLINE NEWS

As answer to our first research question ("*How can a treemap visualization be leveraged to create an interactive control panel for RS?*"), we propose *NewsViz* (see Figure 1). The following sections are organized according to the three central functions of *NewsViz*: 1) provide an overview of the item domain, 2) visualize the user's preference profile, and 3) let this profile interactively be customized.

3.1 Visualize the Domain Space of a News RS

In alignment with common news aggregation websites (e.g. *Google News*[2] and *SmartNews*[3]), the treemap in *NewsViz* is organized as hierarchy of news categories as uppermost hierarchical level and news sources as second hierarchical level. To support users in distinguishing category cells in the treemap easily, each is assigned a specific color. Sources are assigned a background color of their corresponding category but with a different saturation. In this way, cells become distinguishable without cluttering the visualization. To link news articles in the news feed and treemap cells, each article is colored according to its corresponding category-source combination (see Figure 1, in the background).

3.2 Represent the User's Preference Profile

The size (i.e. area) of each treemap cell reflects the number of articles in the news feed that pertain to the corresponding category or source. In the initial setup, all cells on one level have the same size, as the news feed consists of the same number of articles for each category-source combination. This state is also depicted in Figure 1. When preferences of the current user are known beforehand (e.g. elicited using click-through rates), they can be visualized as accordingly adjusted cell sizes in the treemap.

Since assessing the proportional influence of each treemap cell is a central requirement in our approach, the computation of cell alignment follows the algorithm for *squarified treemaps* [3]. Squarified treemaps are an alternative to treemaps that are created by using the *slice-and-dice* strategy. Thereby, squarified treemaps take the aspect ratio of the resulting cells into account thus favoring cells with balanced aspect ratio, i.e. squares. Opposed to slice-and-dice treemaps, squarified treemaps are cognitively easier to interpret, especially in terms of comparing cell sizes.

3.3 Let Users Interactively Adjust Their Preference Profile

While the treemap visualization could already be used to display preference profiles, we also target at supporting users in controlling their personalized news feed. Since preferences in our approach are displayed as cell sizes of the treemap, the most natural interaction concept is to let users directly adjust them. To this end, we follow the interaction concept of *pumping* [1], which is illustrated in Figure 2. In its initial state, cell sizes of the uppermost level in the hierarchy of

[1] http://www.newsmap.jp/
[2] https://news.google.com
[3] https://www.smartnews.com

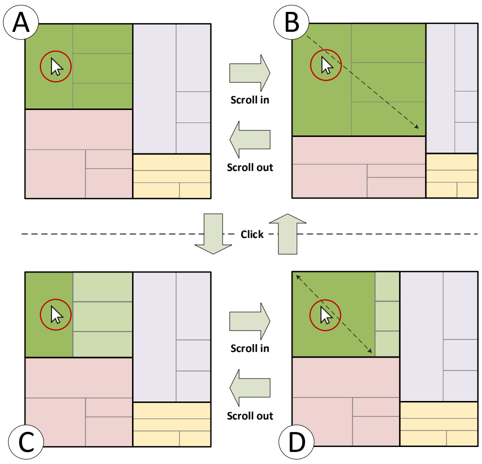


Figure 2: The interaction concept of *NewsViz*. When hovering a treemap cell and scrolling in or out, the size of the corresponding cell increases or decreases, respectively (*A* ↔ *B*). When clicking somewhere in the treemap (*A* ↔ *C* and *B* ↔ *D*), the proportions of cells on the second level can be manipulated in the exact same way (*C* ↔ *D*).

NewsViz can be adjusted. This is done by hovering the mouse cursor over a category cell and moving the mouse wheel. According to the direction in which the wheel is moved, the currently hovered cell enlarges or shrinks, thus mocking a zoom behavior. In this sense, "zooming in" corresponds to enlarging the cell and, as a result, increasing the influence of the corresponding category. Template for this behavior is the concept *zoom and filter* of the *Information Seeking Mantra* [52]. When users click anywhere in the treemap, they can also adjust the size of cells for sources. Adjusting sizes of source cells is done in the very same manner as adjusting category cells. In order to prevent users to lose awareness of sources that are shrunk to a very small size, an according message is shown for each source cell that currently has no influence on the news feed.

Independent of which cells are currently manipulated, each single interaction step triggers an immediate update of the personalized news feed in the background (see Figure 1).

3.4 Implementation Details

We developed *NewsViz* as a web application using the Java framework *Spring Boot* for the backend in tandem with *Vue.js* for the frontend. The treemap visualization is based on the *javascript* library *d3.js* and extends an existing project for web-based treemap visualizations[4]. Background data for news were crawled as preparation for our user study. For this, we used the *NewsAPI*[5] and collected 779 articles from six different news sources, organized in six categories. Sources and categories were selected in order to represent a diverse data sample for the user study. We took special care to choose news sources with different political orientations.

Depending on the proportions of categories and sources in the treemap, the news feed is set up. The entire news feed consists of 100 articles. We chose this number of articles since it contains a reasonable amount of variety and on the other hand is not too large. The influence of each category, divided into sources, determines how many articles are passed to the news feed. Afterwards, articles are sorted according to their recency. This procedure follows one of the typical approaches of how to guarantee recency in news recommendations and is often referred to as *post-filtering strategy* [30]. We decided for a comparable simple recommendation algorithm, since our focus lies entirely on the visualization, the interaction concept, and the user's perception thereof.

4 STUDY SETUP

In order to evaluate *NewsViz* against a baseline system, we conducted an empirical user study. The study was designed as controlled lab study with two conditions: *NewsViz* and a second prototype based on slider widgets.

4.1 Slider-based Prototype

Since we wanted to test the treemap visualization of *NewsViz* isolated from other factors (e.g. the algorithm of news feed composition), the second prototype varies only in replacing the treemap with sliders to indicate preference weights. All other aspects (e.g. hierarchy of sources pertaining to categories and linking articles to categories by color) were maintained identically to the *NewsViz* condition. The result is depicted in Figure 3.

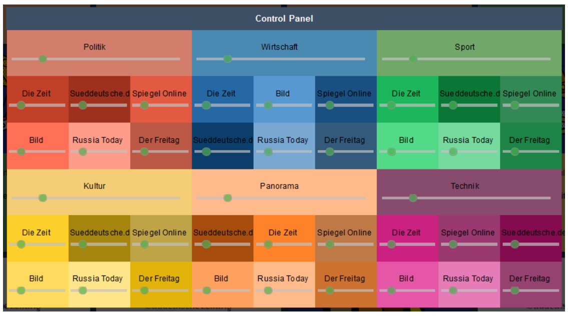


Figure 3: Screenshot of the control panel in the slider-based prototype. Current configuration of categories and sources is the same as in Figure 1.

Consequently, also the behavior on interaction of this baseline prototype mirrors *NewsViz* as closely as possible. Hence, sliders were not independent from each other but instead always displayed their relative value proportionally to the entire distribution of news. As a consequence, when interacting with one slider, the other sliders on the same level (i.e. categories or sources inside one category) reacted in form of adjusting themselves proportionally to yield at any given moment a total of 100% when all sliders are summed up.

4.2 Study Procedure

At the beginning of each user session, the current participant was randomly assigned to one of the two conditions. Then, independent of the condition, participants were asked to fill in a first questionnaire composed of questions regarding their background knowledge and demographical data. The study took place under controlled conditions: participants were alone in a room with a supervisor,

[4] https://github.com/albertopereira/vuejs-treemap
[5] https://newsapi.org/

who was not able to see their computer monitor. Questionnaires were shown using an online tool, which was presented in the same browser window as the prototype (24" screen with a 1920 × 1200px resolution). After finishing the first questionnaire, participants were given a brief introduction to the system (i.e. *NewsViz* or the slider-based variant). It was made clear to them, how to interact with the control panel, that each interaction will trigger an immediate re-calculation of recommendations, and that they could ask questions to their supervisor at any point of the experiment. All participants also received a sheet of paper with a brief paragraph about each news source, in case they were unfamiliar with it. Paragraphs were composed with information from Wikipedia in order to reflect the political orientation of that respective source.

Task 1: After the brief introduction, participants were given the task to use the control panel in order to configure the system regarding their personal preferences of online news. During the task they were allowed to switch between news feed and control panel as often as they wished. Participants were instructed to assess the quality of their recommendations by reading headlines and content teaser of recommended articles. As soon as they think their personalized news feed represents their preferences, participants were requested to fill another questionnaire with questions about their experience during this task.

Task 2: The second task took place as part of the questionnaire. Dependent on their condition, participants were presented with screenshots of four differently pre-configured control panels. Apart from the first profile being shown—which acted for introductory purposes and was neutrally configured—all configurations favored one source over the others. We asked participants to name this most influential source. This task addressed to measure how well participants can spot potential biases.

4.3 Instruments Used

The questionnaire, that was shown first to participants, was composed of questions about their news consumption (e.g. "*How often do you read online news each week?*") and general demographics, e.g. about their age. If not stated otherwise, all items were assessed on 5-point Likert scales.

After participants finished task 1, they were presented with questions about their experience with the system. To measure *transparency* and *overall satisfaction*, we used constructs from the *ResQue* questionnaire [47]. For assessing the individual perceived degree of control, users are able to exert over calculation of the personalized news feed, we used the construct of *interaction adequacy*, which we also took from *ResQue*. For measuring *recommendation quality* and *system effectiveness*, we utilized items that were introduced by Knijnenburg et al. [34]. These instruments were supplemented by three questions, we formulated ourselves for assessing the perceived *overview* of the item space: 1) "*The system helped me to get an overview of the entire spectrum of online news.*", 2) "*The system helped me to get an overview of the entire spectrum of categories.*", and 3) "*The system helped me to get an overview of the entire spectrum of sources.*"

At the end of the questionnaire, participants were given the choice to provide a qualitative comment.

Table 1: Mean values and standard deviations for variables assessed with *NewsViz* and the slider-based version. All variables use 5-point Likert scales. We only found a significant difference (marked with *) in the perceived degree of control (see Section 5.1).

	Treemap Condition		Slider Condition	
Variable	*M*	*SD*	*M*	*SD*
Overview	3.69	0.71	3.60	0.72
Transparency	4.03	1.03	4.29	0.69
Control	4.08^*	0.67	3.45^*	0.98
Recommendation Quality	4.01	0.59	3.97	0.46
System Effectiveness	3.71	0.65	3.71	0.58
Overall Satisfaction	3.94	0.91	3.97	0.84

4.4 Sample

We recruited 63 (31 female) participants for our user study, which were randomly assigned to conditions, resulting in sample sizes of $N = 32$ for the treemap condition and $N = 31$ for the condition based on sliders. The age of our participants ranged from 18 to 52 ($M = 27.02$, $SD = 6.74$) and most of them had a university (49.2 %) or high school degree (36.5 %). As profession, 55.6 % stated being students followed by 41.15 % who were currently working as employees. When asked for habits regarding their news consumption, participants answered that they are somewhat interested in online news ($M = 2.33$, $SD = 1.03$) and mainly did not pay for them during the last year (92.1 %). Participants received no incentive for taking part in our study other than a certificate of participation, which 14 of them needed as requirement for their study program.

5 RESULTS

The conducted experiment was designed to answer our research questions 2 and 3 introduced in Section 1. To answer *RQ2*, questionnaire results are compared between those elicited with *NewsViz* and those elicited with the slider-based prototype. We further uncover relations among results by using a structural equation model (SEM). Finally, to answer *RQ3*, success rates and task completion times of the second task are presented.

5.1 Comparing *NewsViz* to a Slider-based Prototype

Descriptive results of *NewsViz* and the second prototype based on sliders, can be found in Table 1. As can be seen, results for all items are relatively high. For those items that we formulated ourselves to measure *overview*, we calculated Cronbach's alpha in order to assess the internal consistency, which led to an effect of $\alpha = .59$.

In order to test for statistical differences, we compared results for all our six dependent variables between conditions using one-way MANOVA. The multivariate effect with $F(6, 56) = 2.59$, $p = .028$, *Wilk's* $\Lambda = 0.783$, $\eta_p^2 = .217$ for *condition* was statistically significant. The individual dependent variables were subject to ANOVAs in order to assess whether there were any differences of perceived *overview*, *transparency*, *control*, *system effectiveness*, *recommendation quality*, or *overall satisfaction* between conditions. Analyzing the between-subject effects, we could observe that *condition* has

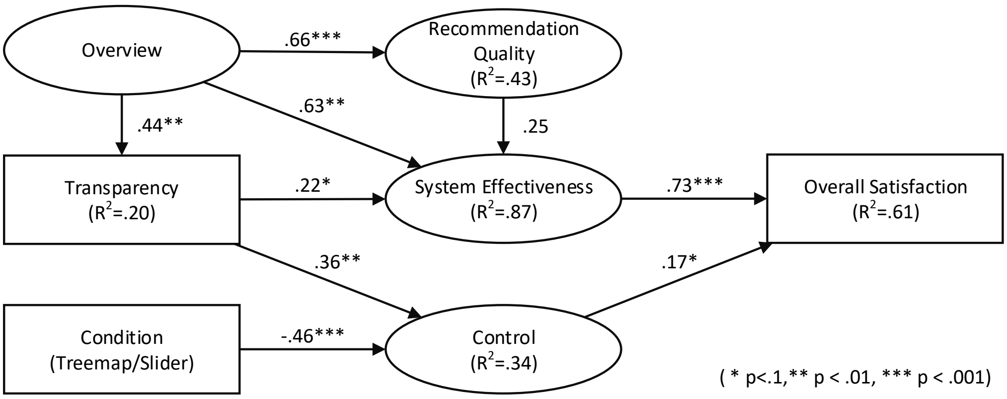

Overview
.66***
Recommendation Quality (R²=.43)
.63**
.44**
.25
Transparency (R²=.20)
.22*
System Effectiveness (R²=.87)
.73***
Overall Satisfaction (R²=.61)
.36**
.17*
Condition (Treemap/Slider)
-.46***
Control (R²=.34)
(* p<.1,** p < .01, *** p < .001)

Figure 4: Structural Equation Model revealing causal dependencies between variables of our experiment. Rectangles represent manifest (observed) variables, while latent (unobserved) constructs are depicted as ellipses. R^2 values are given inside the nodes and denote the explained variance of the corresponding variable. Edges of the graph show standardized parameter weights.

a significant effect on *control*, $F(1,61) = 9.01$, $p = .004$, $\eta_p^2 = .129$. Apparently, *NewsViz* was perceived as easier to control by participants. Other than that, there were no differences with statistical significance.

In order to reveal further dependencies among the aspects measured for both prototypes, we hypothesized a SEM based on the results of our questionnaire (see Section 4.3). The resulting model showed a good fit with the data ($\chi^2(200) = 205.959$, $p = .371$, *CFI* = .988, *TLI* = .986, *RMSEA* = 0.022) and is presented in Figure 4. For setting the SEM up, we used the *R* package *lavaan* [49].

Overview and *condition* acted as exogenous variables, while all other variables of our model were endogenous. *Transparency* and *overall satisfaction* are observed questionnaire items (displayed as rectangles in Figure 4), while *overview*, *recommendation quality*, *system effectiveness*, and *control* are latent composite variables (displayed as ellipses in Figure 4). In order to not overload the graph, we omitted observed manifest questionnaire items for these latent constructs.

One of the most central variables is *overview* as it influences *transparency*, *system effectiveness*, and *recommendation quality*. While not as influential as *overview*, *transparency* shows impact on *system effectiveness* and *control*. *Control* was the only variable that was affected by *condition*. We coded conditions in our experiment with "1" for the condition with *NewsViz* and "2" for the condition with the slider-based variant. In this sense, the negative weight on the edge *condition* ⟶ *control* indicates a higher degree of control in the condition using *NewsViz*, which is in line with our findings using ANOVA stated above. *Control* itself had an effect on *overall satisfaction* indicating that users prefer being in control over their recommendations, though the effect was rather small. Together with *system effectiveness*, *control* was able to explain 61% of the variance of *overall satisfaction*, whereas *system effectiveness* showed the larger extent of. *System effectiveness* was the variable with most entering paths, i.e. it was influenced by other variables the most. It also had the highest amount of explained variance. *Recommendation quality*, as well as *transparency* and *overview* were the predictors for *system effectiveness*. Note, that the regression *recommendation quality* ⟶ *system effectiveness* was not significant ($p = .155$)[6].

5.2 Prevention of Filter Bubbles

In task 2, participants were presented with depictions of different preference profiles and were asked to estimate what the most influential source of the profile is. With this task we address *RQ3* and thus how easy it is to assess biases such as unilateral news consumption and, as a result, how high the risk of filter bubbles is. Screenshots of both conditions for one of the profiles are depicted in Figure 5. The source that objectively had the highest influence in this case was "*Der Freitag*" (45.9%), which was also answered by 100% of participants in the treemap condition and 77.4% in the slider condition. Over all three profiles[7] shown, 87.5% of participants in the treemap condition and 68.8% of participants in the slider condition were able to spot the source with the highest influence. Task 2 took on average 02:16 minutes (*SD*= 00:55) in the treemap condition and 03:08 minutes (*SD*= 01:35) in the slider condition. Multivariate effect for *condition* was statistically significant ($F(2,60) = 10.1$, $p = .000$, *Wilk's* $\Lambda = 0.748$, $\eta_p^2 = .252$). *Condition* had a significant effect on success rate ($F(1,61) = 9.62$, $p = .003$, $\eta_p^2 = .136$), revealing that treemap users could find the most influential source better than users of the slider-based version. The same accounts for task completion times, which were significantly shorter in the treemap condition ($F(1,61) = 6.96$, $p = .011$, $\eta_p^2 = .102$).

6 DISCUSSION

When examining descriptive results of our experiment in Table 1, they confirm that treemaps can effectively be used as interactive input panel for news RS. In qualitative answers, this was further underlined as participants gave statements such as "*Well-grounded,*

[6] We hypothesized this path for logical reasons and thus report it in the SEM, even though it is not significant. We believe, however, that this influence would become stronger in real-world scenarios, where the recommendation quality is more important when assessing the effectiveness of a RS.

[7] Note, that the first profile was shown to introduce the task and thus had no clear most influential source. Consequently, we omitted this profile in analysis of results.

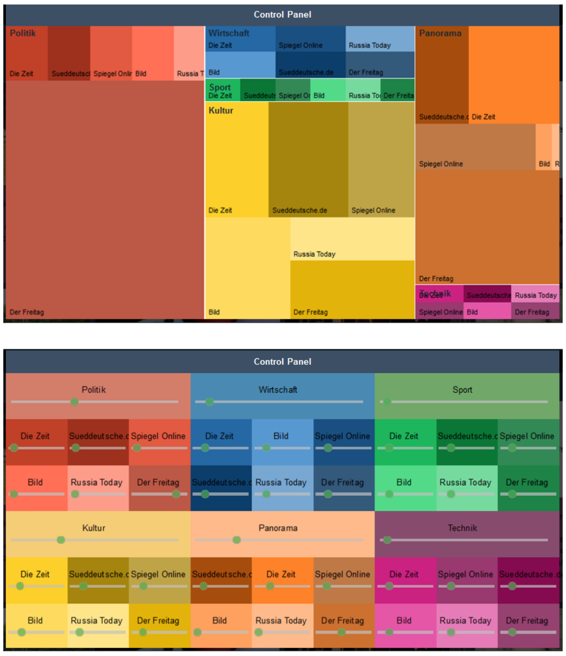


Figure 5: Screenshot of a preference profile that was shown to participants during task 2. The source with the highest influence is in both cases "*Der Freitag*".

interesting system. I absolutely miss something similar on the Internet." (P34) or "*Very innovative and [the system] increases the spectrum of visible online news.*" (P59).

6.1 Comparison of *NewsViz* to a Slider-Based Version

In our second research question, we asked ourselves how *NewsViz* performs in comparison with another system that is not dependent on sophisticated visualizations but uses common slider widgets. Therefore, we deliberately designed the second prototype as equally powerful in terms of interaction modalities. Regarding the results, it became apparent that participants perceived to obtain a higher degree of control with *NewsViz*. We assign this observation to a stronger directness between interaction, visualization and recommendations: Users in the condition with *NewsViz* were able to comprehend more naturally how their interaction with one category or source influenced siblings on the same hierarchical level. When examining sub items of the construct *interaction adequacy*, which we used to measure control, answers most saliently differ for the question "*The system allows me to tell what I like/dislike.*" (*NewsViz*: $M=4.47$, $SD=0.62$; Slider condition: $M=3.65$, $SD=1.17$). Apparently, participants in the slider-based condition did not feel able to express their preferences adequately. In line with prior research [19], we ascribe this to a confusion due to an overcrowded interface. This assumption is also backed up by qualitative statements of participants, which, for instance, experienced the system as "*confusing and irritating*" (P49).

Another reason for this confusion may originate from the slider behavior: When one slider position was changed, the other sliders adjusted themselves to result in an overall distribution that sums up to 100%. Even though adjusting the size of treemap cells also affected the other cells on the same level in *NewsViz*, we assume that this behavior was perceived as more natural. The total size of the treemap gives users a point of reference, making it thus more comprehensible that cells have to arrange themselves in order to fit into the given frame.

Apart from the perceived degree of control, we did not find any statistically significant differences in the results. This is especially noteworthy in terms of the perceived degree of overview. While this indicates that overview in RS is not necessarily tied to complex visualizations, we assume that some potential for overview in RS was left unused in our approach. When comparing screenshots of both prototypes (see Figure 1, Figure 3 and Figure 5), it can be seen that the slider variant is already rather crowded, while *NewsViz* still appears comparable tidy. This is in line with aforementioned prior work about slider elements [19].

Results of task 2 show that participants of the treemap condition were able to spot the most influential source more easily. This became visible through success rates and task completion times (see Section 5.2), which were both significantly better for the treemap condition. As a consequence, we deem treemaps to be superior in conveying a natural sense of the entirety of news items, in relating proportions of own preferences to the rest of the space, and thus in raising awareness for possible biases due to overrepresented sources. This could especially become relevant when users are presented with preference profiles elicited implicitly in the past. With help of the treemap visualization they could rapidly apprehend their preference model, spotting possible biases and adjusting it to resolve these biases or regarding current preferences. We, though, note that users could still create profiles that neglect inconvenient sources and thus self-reliantly creating a filter bubble. Yet, our approach would make it harder to become unaware of such sources.

6.2 Dependencies Among Quality Aspects of RS

In Section 5.1, we introduced a SEM in order to uncover otherwise hidden relations between the different quality aspects of RS that we elicited in our study. As this model shows, *overview* was the most influential variable, influencing *transparency* as well as *system effectiveness* and *recommendation quality*. The influence over the path *overview* → *transparency* appears rather natural: when users perceive a sense of overview of the underlying item space, they understand the data used for recommending better and thus perceive the entire RS as more transparent. While the opposite relation is imaginable too (i.e. degree of transparency accounts for the perception of overview), our understanding of overview follows basic literature on information visualization that treats the notion of overview as *awareness* of an item space [26], which happens *first* [52] in human-computer-interaction. Our self-invented items reflect this understanding (see Section 4.3).

Not all influences emerging from *overview* are that easy to interpret, though. The path *overview* → *recommendation quality* reveals a direct influence of the perceived degree of overview on the quality of recommendations. Note, that this regression even yields a rather high amount of explained variance in *recommendation quality* (43%). That we found a causal relation between such ostensibly isolated variables, emphasizes the complexity of measuring the quality of recommendations with user studies. Recently, we made a similar observation [36] and found that the perceived quality of a recommendation is significantly influenced by how well the system explains its reasons for recommending this particular item. Apparently, users take a lot more aspects into account when assessing the quality of a system's recommendations than solely the suggested items.

Besides *transparency* and *recommendation quality*, *system effectiveness* was influenced by *overview* as well. Deconstruction of the composite variable *system effectiveness* in its latent components can yield some insights into the rationale of this regression. Especially two items help to understand the influence of *overview* on *system effectiveness*: "*The system makes me more aware of my choice options.*" and "*I make better choices with the system.*" In this light, experience with our RS appears to form an arc that begins with a general overview, which then leads over awareness of choice options, to better decisions, and finally to a general perception of system effectiveness. This sequence can also be found in general literature on decision making, e.g. in [53].

The regression *transparency* ⟶ *control* also underlines that variables in user studies should not be treated independently of each other. Apparently, control in RS is positively influenced by how transparent the system lays out its inner workings and reasons for recommending to the user. We think this causal relation is naturally comprehensible: to control a complex system (such as a RS), a user needs to some degree understand how the system works and thus where or how to interact with it. Our observation here adds to evidence found by other authors who also underline the importance of transparency to foster control in RS [22, 58].

6.3 Limitations

There were also some limitations to our experiment. First, the sample is comparatively small—especially in terms of performing structural equation modeling. As a result, the model lacks robustness to some degree and interpretations need to be made with care. However, since some of our deductions are not only relying on the SEM but also on comparative analysis and prior observations, we are confident that the trends in our model are valid and would become even stronger with more data. We also acknowledge that further studies with a more representative sample (e.g. in terms of education) should be conducted.

Second, the value for Cronbach's alpha calculated for our construct to measure overview, indicates that there are some inconsistencies within the answers. Nonetheless, we believe that the wording of items is rather clear and that the assessment of overview is reliable.

A third limitation pertains to the number of cells currently displayed in *NewsViz*. In our experiment both prototypes used six categories and six sources, resulting in a total of 36 cells. When applied to real news aggregation portals, the number of categories and sources would probably be higher than this. Also the hierarchy of categories would be more sophisticated (e.g. splitting *politics* into *local* and *global*). While our experiment shows that the slider-based variant was already experienced as cluttered and confusing, we deem the treemap visualization to yield potential for adding several further categories and sources (see, for instance, the large number of articles displayed in *Newsmap*[8]). Nonetheless, the treemap visualization would at one point become overcrowded too. To encounter this, only a small part of the entire profile could be visualized at a given time. When combined with *overview+detail* visualizations [7] (e.g. in form of a minimap) users would still be able to keep awareness of their entire profile.

7 CONCLUSIONS AND OUTLOOK

In this paper, we introduce a novel interface for controlling news recommender systems. The resulting system, *NewsViz*, utilizes a treemap to display the domain space of news as hierarchy composed of news categories and sources. By adjusting cell sizes of this treemap, users can interactively control the influence of the corresponding category or source, respectively. In an empirical user study, *NewsViz* was compared to a baseline system using slider widgets for interaction. Results indicate that *NewsViz* scored as good as the slider-based system in dimensions such as perceived overview, transparency and user satisfaction. The degree of control, however, was perceived as higher for *NewsViz*. By applying structural equation modeling to our results, we were able to identify additional interesting causal influences of overview on transparency, system effectiveness and recommendation quality.

This observation underlines how complex quality assessments of recommendations are and we conclude that more research is necessary about relations between quality aspects of recommender systems. User study results that at first seem isolated should be put into context (e.g. by structural equation modeling) before making reliable assertions about their meaning. In this sense, our structural equation model reveals that the degree of *overview* significantly influenced other quality aspects of the system. Yet, how much overview of the item space is perceived by users is rarely used as means for assessing a system's quality. In future research we thus plan to investigate this aspect in greater depth. Thus far, for instance, a reliable instrument for measuring the perceived degree of overview is missing.

In the future we also plan to test *NewsViz* in other situations. At the moment, for instance, users always start with a neutral treemap, corresponding to cold start situations. In upcoming research, however, we plan to record implicit user feedback, e.g. by logging clicks and dwell times, and visualize the resulting profiles to users. Their reactions could be insightful in two ways: 1) how users react, when they are presented with their implicitly recorded behavior; and 2) whether they will appreciate it when their otherwise hidden profile is not only shown but also made controllable to them.

We further believe that attributes of treemaps can be used more extensively in *NewsViz*. For instance, the inherent ability to depict several levels of hierarchy bears so far unused potential. In our setting, we depicted two hierarchical levels. In practical settings, however, especially categories are typically organized in more than one categorical layer. With such an advanced interface we also plan further user surveys, in which we will also compare *NewsViz* to other baseline systems (e.g. even simpler ones with less interaction modalities).

[8] http://www.newsmap.jp/

REFERENCES

[1] Toshiyuki Asahi, David Turo, and Ben Shneiderman. 1995. Using Treemaps to Visualize the Analytic Hierarchy Process. *Information Systems Research* 6, 4 (1995), 357–375. https://doi.org/10.1287/isre.6.4.357

[2] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. https://doi.org/10.1126/science.aaa1160

[3] Mark Bruls, Kees Huizing, and Jarke J. van Wijk. 2000. Squarified Treemaps. In *Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Amsterdam (Data Visualizaion 2000)*. Springer Vienna, 33–42. https://doi.org/10.1007/978-3-7091-6783-0_4

[4] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for Recommender Systems: The Past, the Present and the Future. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 123–126. https://doi.org/10.1145/2959100.2959158

[5] Bruno Cardoso, Peter Brusilovsky, and Katrien Verbert. 2017. IntersectionExplorer: the flexibility of multiple perspectives. In *Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2017) (IntRS 2017)*. 16-19

[6] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and Capturing Complex User Interests for Exploratory Search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, 498–509. https://doi.org/10.1145/3301275.3302321

[7] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. 2009. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces. *Comput. Surveys* 41, 1 (2009), 2:1–2:31. https://doi.org/10.1145/1456650.1456652

[8] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America* 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[9] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (2016), 56–62. https://doi.org/10.1145/2844110

[10] Dominic DiFranzo and Kristine Gloria-Garcia. 2017. Filter Bubbles and Fake News. *XRDS* 23, 3 (2017), 32–35. https://doi.org/10.1145/3055153

[11] Michael D. Ekstrand, Daniel Kluver, F. Maxwell Harper, and Joseph A. Konstan. 2015. Letting Users Choose Recommender Algorithms: An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. ACM, 11–18. https://doi.org/10.1145/2792838.2800195

[12] Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations Through Listening to Users. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 221–224. https://doi.org/10.1145/2959100.2959179

[13] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion Space: A Scalable Tool for Browsing Online Comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 1175–1184. https://doi.org/10.1145/1753326.1753502

[14] Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80, S1 (2016), 298–320. https://doi.org/10.1093/poq/nfw006

[15] Emden Gansner, Yifan Hu, Stephen Kobourov, and 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, 345–348. https://doi.org/10.1145/1639714.1639784

[16] Emden R. Gansner, Yifan Hu, and Stephen G. Kobourov. 2014. Viewing Abstract Data as Maps. In *Handbook of Human Centric Visualization*, Weidong Huang (Ed.). Springer New York, New York, NY, 63–89. https://doi.org/10.1007/978-1-4614-7485-2_3

[17] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382. https://doi.org/10.1016/j.ijhcs.2013.12.007

[18] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the Filter Bubble? *Digital Journalism* 6, 3 (2018), 330–343. https://doi.org/10.1080/21670811.2017.1338145

[19] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. 2019. Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 69–77. https://doi.org/10.1145/3298689.3347014

[20] Jaron Harambam, Natali Helberger, and Joris van Hoboken. 2018. Democratizing algorithmic news recommenders: How to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180088. https://doi.org/10.1098/rsta.2018.0088

[21] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting Users in Control of Their Recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. ACM, 3–10. https://doi.org/10.1145/2792838.2800179

[22] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27. https://doi.org/10.1016/j.eswa.2016.02.013

[23] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

[24] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

[25] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, 241–250. https://doi.org/10.1145/358916.358995

[26] Kasper Hornbæk and Morten Hertzum. 2011. The notion of overview in information visualization. *International Journal of Human-Computer Studies* 69, 7 (2011), 509–525. https://doi.org/10.1016/j.ijhcs.2011.02.007

[27] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 7–10. https://doi.org/10.1145/2959100.2959186

[28] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders–Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10:1–10:46. https://doi.org/10.1145/3001837

[29] Martijn Kagie, Michiel van Wezel, and Patrick J.F. Groenen. 2011. Map Based Visualization of Product Catalogs. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, Boston, MA, 547–576. https://doi.org/10.1007/978-0-387-85820-3_17

[30] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems – Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

[31] Rahul Katarya, Ivy Jain, and Hitesh Hasija. 2014. An interactive interface for instilling trust and providing diverse recommendations. In *2014 International Conference on Computer and Communication Technology (ICCCT)*. 17–22. https://doi.org/10.1109/ICCCT.2014.7001463

[32] Mohammad Khoshneshin and Nick W. Street. 2010. Collaborative Filtering via Euclidean Embedding. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10)*. ACM, 87–94. https://doi.org/10.1145/1864708.1864728

[33] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender Systems for Self-Actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 11–14. https://doi.org/10.1145/2959100.2959189

[34] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[35] Joseph A. Konstan and 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

[36] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–12. https://doi.org/10.1145/3290605.3300717

[37] Johannes Kunkel, Benedikt Loepp, and 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, 3–15. https://doi.org/10.1145/3025171.3025189

[38] Jaron Lanier. 2011. *You are not a gadget: a manifesto*. Vintage Books, New York, NY, USA.

[39] Yu Liang. 2019. Recommender System for Developing New Preferences and Goals. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 611–615. https://doi.org/10.1145/3298689.3347054

[40] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, 1097–1101. https://doi.org/10.1145/1125451.1125659

[41] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

[42] Sayooran Nagulendra and Julita Vassileva. 2016. Providing awareness, explanation and control of personalized filtering in a social networking site. *Information Systems Frontiers* 18, 1 (2016), 145–158. https://doi.org/10.1007/s10796-015-9577-y

[43] Philip M. Napoli. 2015. Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy* 39, 9 (2015), 751–760. https://doi.org/10.1016/j.telpol.2014.12.003

[44] Efrat Nechushtai and Seth C. Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (2019), 298–307. https://doi.org/h10.1016/j.chb.2018.07.043

[45] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* The Penguin Press, New York, NY, USA.

[46] Denis Parra, Peter Brusilovsky, and Christoph Trattner. 2014. See what you want to see: visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces (IUI '14)*. 235–240. https://doi.org/10.1145/2557500.2557542

[47] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, 157–164. https://doi.org/10.1145/2043932.2043962

[48] Christian Richthammer and Günther Pernul. 2017. Explorative Analysis of Recommendations Through Interactive Visualization. In *E-Commerce and Web Technologies (EC-Web 2017)*. Springer International Publishing, 46–57. https://doi.org/10.1007/978-3-319-53676-7_4

[49] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012). https://doi.org/10.18637/jss.v048.i02

[50] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking Spatial Visualization Through Semantic Enhancement. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, 179–190. https://doi.org/10.1145/3025171.3025233

[51] Ben Shneiderman. 1992. Tree Visualization with Tree-Maps: 2-d Space-Filling Approach. *ACM Transactions on Graphics* 11, 1 (1992), 92–99. https://doi.org/10.1145/102377.115768

[52] Ben Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE, 336–343. https://doi.org/10.1109/VL.1996.545307

[53] Allan D. Shocker, Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi. 1991. Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters* 2, 3 (1991), 181–197. https://doi.org/10.1007/BF00554125

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

[55] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the Unknown: Visualising Consumption Blind-spots in Recommender Systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC '18)*. ACM, 1396–1399. https://doi.org/10.1145/3167132.3167419

[56] Daniel Trielli and Nicholas Diakopoulos. 2019. Search As News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 453:1–453:15. https://doi.org/10.1145/3290605.3300683

[57] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, 239–250. https://doi.org/10.1145/3172944.3172959

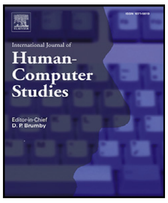
[58] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining Recommendations in an Interactive Hybrid Social Recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, 391–396. https://doi.org/10.1145/3301275.3302318

The following article was published in the International Journal of Human-Computer Studies, Volume 172, Johannes Kunkel and Jürgen Ziegler, A comparative study of item space visualizations for recommender systems, Copyright Elsevier (2022).

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs



A comparative study of item space visualizations for recommender systems

Johannes Kunkel *, Jürgen Ziegler

Interactive Systems Research Group, University of Duisburg–Essen, Forsthausweg 2, 47057 Duisburg, Germany

ARTICLE INFO

Dataset link: https://data.mendeley.com/datasets/hw39d5kxs6/3

Keywords:
Recommender systems
Information visualization
Empirical user studies
User experience
Maps
Treemaps

ABSTRACT

Recommender systems aim at supporting users in their search and decision making process by selecting a small number of likely relevant items from a large set of options. Although automatically filtering unmanageably large item sets down to a few recommendations often produces results that match the user's interests well, it also prevents users from understanding and exploring items in their larger context. This may reduce users' perception of transparency and controllability of the system. Visualizations have been proposed as a means for overcoming this problem, with some visualizations providing a complete overview of the entire space of available items. However, thus far item space visualizations have rarely been investigated and compared in user studies. To address this, we developed and empirically compared three applications that present the user with personalized music recommendations embedded in a visualization of the entire item space. The three applications display the same item space as a list, as a treemap, and as a map, respectively. We compared these applications in an online user study and found, against our expectations, that they did not differ much in how the recommendations are perceived. Perception of transparency, recommendation quality, and degree of control over the recommendations received relatively high scores over all three applications. However, we did find a difference in hedonic user experience and perceived novelty of the recommendations. Both factors were perceived to be higher in the map condition. Backed up by a mediation analysis, we argue that a *halo* effect is the reason for the observed perceived novelty: participants transferred the novelty of the application to the novelty of the recommendations.

1. Introduction

Recommender systems (RS) are established tools that proactively filter large item spaces—i.e. the entire set of available items in web stores, music streaming portals, social networks, and other digital platforms—to suggest only small personalized sets of items that match their users' preferences. RS have been found to be able to increase sales and user conversion rates, to promote customer loyalty and retention, and to improve overall user satisfaction (Ricci et al., 2022). Over the last decades, recommendation algorithms have achieved high accuracy in predicting a user's tastes based on their prior interactions with the system (Gunawardana and Shani, 2015).

However, RS typically present only a small filtered set of items, but do not support exploration of larger proportions or even the entirety of the underlying item space. This introduces a range of potential issues: Users may not be aware that they are being presented with filtered content, and thus become trapped in *filter bubbles* (Pariser, 2011). Confidence in choices may decline because users do not understand the item space and thus their range of options (Jannach and Adomavicius, 2016). Transparency of the RS may also be negatively affected when users have no means of assessing the relevance of its output, i.e. the recommended items, in the context of other items that are not recommended or their own preferences (Hellman et al., 2022; Tintarev and Masthoff, 2022). Users might also regret their choices because they feel they missed opportunities (Iyengar and Lepper, 2000). Presenting a limited subset of items could also prevent users from exploring and developing their personal preferences (Knijnenburg et al., 2016; Petridis et al., 2022), and even pose a threat to human creativity in general (Knijnenburg et al., 2016). Fairness and ethical issues may arise as well (Dandekar et al., 2013; Burke et al., 2018). Finally, in *cold start situations*, where a new user approaches a RS, the system is simply unable to suggest appropriate items to them Ricci et al. (2022).

As an alternative to presenting a limited list of recommendations, various graphical interfaces have been proposed to support users in exploring the entire item space. To implement these interfaces, various techniques from the field of *information visualization* (InfoVis) have been applied. They range from rather simple chord diagrams (Tintarev et al., 2018) to map-based visualizations (Sen et al., 2017; Knees et al., 2019). An early example is *TVLand* (Gansner et al., 2010). All items of a TV show dataset are distributed on a two-dimensional plane, so that the distance between item positions represents how similar they are.

[*] Corresponding author.
E-mail address: johannes.kunkel@uni-due.de (J. Kunkel).

https://doi.org/10.1016/j.ijhcs.2022.102987
Received 19 April 2022; Received in revised form 29 November 2022; Accepted 22 December 2022

In addition, the map indicates areas of high or low interest to the active user by different colors. In previous research, we presented a similar application that displays the item space of movies as a three-dimensional landscape in which elevations and recesses indicate areas of high and low user preferences in a latent factor space (Kunkel et al., 2017). A similar combination of item space visualization and personalized recommendations has also been applied to the domain of music (Andjelkovic et al., 2019) and university courses (Ma et al., 2021).

Since there are several ways to display recommendations in the context of the entire item space, there are also various possible implications for how they are perceived by users—e.g. as how transparent, controllable, or satisfying users perceive recommendations to be. In particular, we formulated the following research questions:

RQ1: How does the type of item space visualization influence users' perception and understanding of the item domain and the recommendations?

RQ2: How does the type of item space visualization influence how users interact with the visualization and the recommendations?

We acknowledge that the effectiveness of using visualizations in RS may depend on individual user characteristics (e.g. see Millecamp et al., 2018; Jin et al., 2019). For this reason, we raise a third research question:

RQ3: How do users' characteristics influence how they perceive, understand, and interact with the item space and the recommendations?

To our knowledge, there are no user studies so far that compare different ways of visualizing large item spaces in RS and can answer the RQs above. For this reason, we developed three web applications that visualize the entire item space of a music dataset. All three prototypes allow browsing the item space, rating and listening to songs, and receiving song recommendations, but use different types of visualization: a list, a treemap, and a map.

Based on these prototypes, we conducted an empirical online user study.[1] Contrary to expectations based on previous work, we found that the prototypes did not differ significantly in terms of the overview of the item domain that users obtained, nor in terms of how transparent or controllable the recommendations were perceived. However, we observed that participants in the map condition experienced the highest *hedonic user experience* and perceived their recommendations as more novel compared to participants who were presented with a list or a treemap. To uncover the reasons for this, we conducted a mediation analysis and found that the relationship between visualization type and the perceived novelty of recommendations was fully mediated by the hedonic user experience. In other words, the novelty of recommendations perceived by users is due to the general perception of the application as novel, rather than a factual recommendation novelty.

The remainder of this article is structured as follows: In Section 2, we present a thorough literature review of the ongoing research of InfoVis in RS with special focus on visualizing entire item domains. Based on the gaps identified in existing research, we formulated the goals of our research in Section 3. This is followed by Section 4, where we lay out the structure and creation process of the dataset that our prototypes are based on. These prototypes are described in Section 5. In Section 6 we introduce the design and the experimental procedure of our user study; its results are presented in Section 7. Finally, we discuss the results in Section 8 and provide suggestions for future work in Section 9.

2. Related work

The abundance of items in modern repositories such as web stores, news portals, social media sites, and streaming platforms makes it increasingly difficult for users to find content they like. To support users in exploring the item domain, a number of approaches have been taken. Active customization of filter settings by users, for instance, is a method that is especially suitable for exploring domains of *search products*. Items in this domain typically have well-defined attributes that users can search for (e.g. searching for a laptop with a certain CPU speed). The item range of *experience products*, on the other hand, is not as easy to explore using filters. Due to the complexity of evaluating experience products prior to purchasing, users tend to have a more vague search goal and tend to take inspiration from others in making their decision, such as salespersons or automated RS (Xiao and Benbasat, 2007). Examples of such experience products are perfumes, movies, or clothing.

2.1. Recommender systems

Based on a user's click history, purchases, or consumptions, RS proactively suggest a set of similar or matching items to that user. Among other benefits, RS have the potential to increase purchases, user satisfaction, and customer loyalty (Ricci et al., 2022). A popular domain where RS play a key role is music, for example in Spotify,[2] Apple Music,[3] or Deezer.[4] One reason why this domain particularly benefits from recommendations is that music search is mostly undirected (Cunningham et al., 2003). Therefore, exploration is a central task that music RS should support (Schedl, 2017).

In industry, a common approach of creating music recommendations (e.g. reported in the context of Spotify (McInerney et al., 2018) and Deezer (Bendada et al., 2020)) is based on the *multi-armed bandit* optimization problem (Anantharam et al., 1987). At the core of this method is the metaphor of a gambler playing a series of slot machines. Belonging to the family of reinforcement learning problems, the multi-armed bandit problem's key aspect is to find a trade-off between *exploration* (gathering information for algorithmic optimization) and *exploitation* (using the previously gathered information to maximize the outcome). Applied to a music RS, the problem translates to suggesting items to a user that best match their preferences (exploitation), while also learning new preferences of the active user (exploration). Another approach was presented by Anderson et al. (2020) and is based on the *word2vec* embedding algorithm (Mikolov et al., 2013). The algorithm, typically used on words in documents, is applied to songs in playlists: songs, which frequently occur close to each other in the same playlists, are also embedded close to each other in the resulting vector space. To find recommendations for a particular seed song, the algorithm selects the most similar songs based on the smallest cosine similarity to that seed item. Apart from that, many other aspects can be considered in real-world settings. In an experiment using bots to reverse engineer the Spotify algorithm, Eriksson et al. (2019) found that recommendations are sometimes influenced by demographic data. This is supported by statements of McInerney et al. (2018), who list different aspects of contextual factors, such as time of day, that may be taken into account when calculating recommendations at Spotify.

In the research and development of RS, quality criteria that go *beyond* accuracy have long been neglected (McNee et al., 2006; Konstan and Riedl, 2012). Yet, such user-centric aspects may influence the overall perceived quality of a RS as well (Pu et al., 2011; Kunkel et al., 2019). Contemporary RS often still appear as *black boxes* to their users, even though many approaches have been presented on

[1] Supplementary material of this study can be found at https://data.mendeley.com/datasets/hw39d5kxs6/3.

[2] https://open.spotify.com/

[3] https://www.apple.com/apple-music/

[4] https://www.deezer.com

how to increase system transparency (Herlocker et al., 2000; Sinha and Swearingen, 2002; Zhang and Chen, 2020). Besides other effects, it has been observed that increasing the transparency of a RS can improve the perceived quality of recommendations (Kunkel et al., 2019), their acceptance (Herlocker et al., 2000; Giboney et al., 2015), and users' confidence in recommendations (Sinha and Swearingen, 2002). The transparency of a RS is also closely related to how well users are able to control recommendations and interact with the system (Tintarev and Masthoff, 2015; He et al., 2016; Kunkel et al., 2020). Enabling users to exert control over their recommendations can increase their satisfaction with the system (Roy et al., 2019) and also increase perceived recommendation accuracy (O'Donovan et al., 2008). Other user-centered quality criteria include the novelty of recommendations (Herlocker et al., 2004; Hurley and Zhang, 2011; Vargas and Castells, 2011). If users are aware of an item already, it rarely constitutes a good recommendation—even if it accurately matches the user's preferences.

When users rely too much on recommendations and are unaware of other parts of the information space, they can become trapped in *filter bubbles* (Pariser, 2011). While some authors consider such effects to be exaggerated (Haim et al., 2018) or primarily brought about voluntarily by users themselves (Bakshy et al., 2015), others link them to various negative individual or even societal consequences (Del Vicario et al., 2016; Flaxman et al., 2016; DiFranzo and Gloria-Garcia, 2017). To encounter such potential effects, a system should help users stay aware of the entire item space (Nagulendra and Vassileva, 2014), which can also increase other positive attributes of RS. For example, it has been shown that the diversity of consumed products increases when users are able to explore the item space (Andjelkovic et al., 2019), which can help them make more educated decisions (Tintarev et al., 2018). When users are able to understand the item space, and thus the range of options available to them, it can improve their confidence in their choices (Jannach and Adomavicius, 2016) and reduce their fear of missing out (Iyengar and Lepper, 2000). To this, Knijnenburg et al. (2016) add the concept of *self-actualization*: users should be provided with means to understand their own preferences, but also to explore possible alternatives so that they can develop new preferences over time. Dandekar et al. (2013) point out that overly personalized RS may also lead to a societal polarization, while Burke et al. (2018) raise issues of fairness that may run contrary to personalization. Finally, from a more practical perspective, users need ways to manually browse the items of a system in *cold start situations*, i.e. when the RS has not yet collected enough data about the user's preferences in order to generate personalized recommendations (Ricci et al., 2022).

2.2. Item space visualizations in recommender systems

One field where the development of tools to efficiently browse and explore large amounts of data plays a key role is InfoVis. In their seminal work on InfoVis, Card et al. (1999) frame the purpose of InfoVis as "to amplify cognition". Users of such systems realize this amplification by projecting an *external* representation (i.e. the InfoVis system) onto an *internal* representation (i.e. their mental model of the system). A well-designed InfoVis application thus reduces the cognitive workload, which is especially important when dealing with large information domains. Numerous design guidelines and interaction patterns have been formulated to support the development of InfoVis applications. Probably the most prominent among them is the *information seeking mantra* by Shneiderman (1996): overview first, zoom and filter, then details-on-demand.

InfoVis methods have also been applied to the domain of RS (e.g. Parra et al., 2014; Du et al., 2018; Tsai and Brusilovsky, 2019a). For example, perceived transparency and acceptance of recommendations can be increased by using bar charts (Du et al., 2018). By using chord diagrams, Tintarev et al. (2018) showed that such visualizations can help raise user awareness of potential blind spots regarding the item space, thus targeting the issues discussed above. Cardoso et al. (2019) introduced *IntersectionExplorer*, a dashboard designed to make the influence of different RS in a hybrid setting more understandable. A user study confirmed that *IntersectionExplorer* is easy to use and can help find more relevant items. In their system *SmallWorlds*, Gretarsson et al. (2010) used a graph-based approach to visualize the influence of recommendations and let users interactively control this setting. As a result, users perceived high levels of satisfaction and control.

One way in InfoVis to efficiently display large hierarchical tree structures on two-dimensional screens are *treemaps* (Shneiderman, 1992). Since such hierarchical structures are commonly subject of visualizations in RS, treemaps have also been widely used in this context. They have, for instance, been utilized to display the recommendation space (i.e. the set of personalized recommendations for a given user), where cell sizes indicate the fit of each recommendation to the current user's preference profile (Katarya et al., 2014; Richthammer and Pernul, 2017). In a past experiment, we used a treemap to display the otherwise hidden user profile (Kunkel et al., 2020). In this system, *NewsViz*, user profiles are presented within the entire item space of news articles. With *NewsViz*, we showed that this can help make users aware of the composition of their profile and whether it is imbalanced and thus prone to bias. Something similar has also been observed by Torrens and Arcos (2004), who demonstrate the ability of treemaps to provide users with an overview of a music library. Finally, Chang et al. (2019) used a treemap to visualize and manipulate search queries, which motivated users to interact and led to higher user satisfaction.

While treemaps are well-established means of visualizing hierarchical data structures and comparing quantitative variables within them, other forms of InfoVis have been found to be able to convey similarities between data items better (Biuk-Aghai et al., 2017). For this, a region metaphor (Fabrikant et al., 2006), which represents similarities as distances in a map-like visualization, seems more appropriate. This *spatialization* (Kuhn and Blumenthal, 1996; Skupin and Fabrikant, 2003) takes advantage of the fact that using a spatial form of thinking is an innate human trait (Gärdenfors, 2004). Spatial distances have proven to be intuitively understood to represent the relatedness of markers within a map, with smaller distances representing higher relatedness, also known as *"first law of geography"* (Tobler, 1970). Interestingly, the first law of geography was found to apply to depictions of non-spatial data as well (Montello et al., 2003). Accordingly, many InfoVis applications have been developed that represent large product spaces as artificial maps (e.g. Pampalk et al., 2002; Sen et al., 2017; Meinecke et al., 2022).

An early example of spatial maps, *nepTune* (Knees et al., 2006), is placed in the domain of music—a domain that particularly benefits from exploration interfaces. *nepTune* displays the item space of songs as a three-dimensional geographic landscape that can be explored by users. In a user study with *nepTune*, participants responded very positively to the application. We reached a similar conclusion, in a user study with a system that displays the item space of movies as a three-dimensional landscape (Kunkel et al., 2017). As part of this application, users can use metaphorical tools to model the surface of this landscape and shape hills and valleys to express areas of high and low preference, respectively. The user study showed that users perceived the system as transparent and enjoyed using it. Sen et al. (2017) use the map metaphor to visualize the item space of different domains based on the knowledge stored in Wikipedia. To do so, they leverage connections between articles to project any item space onto a two dimensional embedding—Assuming the items can be found in Wikipedia. Petridis et al. (2022) found that personalizing the interface in visualized music recommendations helps users *anchor* their personal music tastes in the visualization. While Petridis et al. use a graph-based approach, Liang and Willemsen (2021) represent the neighborhood of the user's music taste as *contour plot*. This helped users discover new music genres and was experienced as more comprehensible in contrast to a baseline system with bar charts. In another approach, Andjelkovic et al. (2019) present *MoodPlay*, a personalized music interface. In *MoodPlay*, the current user profile is visualized as point in a map of music tracks. By listening to these music tracks, the point moves in the space and a trail indicates the active user's listening history. As one of the few existing examples of using comparative studies to evaluate item space visualizations in RS, Andjelkovic et al. tested their system in a user study where they compared it to a baseline without item space visualization. As a result, they found that participants perceived the interactive map as more transparent and easier to control. Recently, Ma et al. (2021) presented *CourseQ*, a visual tool that recommends university courses to students based on a map metaphor. They also compared their application to a baseline without map visualization and found that *CourseQ* was able to increase the perceived transparency of recommendations as well as their acceptance.

2.3. User characteristics in recommender systems and information visualization

User preferences for the structure of music taxonomies (i.e. whether music should be organized by mood, genre, or activity) were found to be related to the *Big Five* personality traits (McCrae and John, 1992; Ferwerda et al., 2015). In particular, Ferwerda et al. observed that participants high in *openness* were more likely to browse music by mood and participants high in *conscientiousness* were more likely to browse music by activity. *Neuroticism* correlated positively with preference for an activity-based and genre-based music organization. Similarly, the Big Five also correlated with musical tastes (Rawlings and Ciancarelli, 1997). High values for *extraversion*, for instance, relate to taste for pop music. Other correlations with the Big Five were also observed in relation to other aspects in the evaluation of user-centered research in RS (Tkalcic and Chen, 2015). Millecamp et al. (2020) further showed that affinity for explanations in RS can also depend on personality traits; more specifically, Millecamp et al. found that users with high *musical sophistication* and low *openness* used explanations more often.

Especially in terms of how users interact with more complex visual interfaces, *visual memory* was found to be an important predictor (Millecamp et al., 2018; Jin et al., 2019). Conati et al. (2014) found that task performance when using an InfoVis interface was dependent on visual memory capacity. It was also found to influence preference for vertical or horizontal orientation of the graphical interface. In another experiment, Ziemkiewicz et al. (2011) observed that the *locus of control* (i.e. whether a person feels in control over external events in their life) influences users' preference for an indented or a contained representation of a tree structure. In addition to locus of control, *information literacy* (Boy et al., 2014), which describes how well users can handle data visualizations, was found to influence perceptions of intelligent systems that use InfoVis components (Lallé and Conati, 2019). In the domain of music, it was observed that the *musical sophistication* (Müllensiefen et al., 2014) can influence how users interact with an intelligent system (Millecamp et al., 2018).

3. Research goals

In summary, prior research shows that InfoVis methods are capable of visualizing large item spaces in RS and thereby improving user-centered features such as the perceived transparency, diversity, and control of recommendations. These methods thus can alleviate problems that may arise when only a small set of recommendations is presented and users cannot explore the underlying item space appropriately. Little research has been conducted on how this underlying item space should be visualized and how different types of visualization affect the perception of a RS, though.

Designing InfoVis systems that employ recommendations and with which user studies are conducted to gain insights into perceptional user variables is not trivial and can be performed in a variety of ways. Of the empirical studies that use systems that visualize the entire item space and embed a RS in it, few have done so in a comparative manner (e.g. Cardoso et al., 2019; Andjelkovic et al., 2019; Ma et al., 2021). Others do not attempt to represent the item space in its entirety (e.g. Chen and Tsoi, 2011; Parra et al., 2014; Chang et al., 2019; Liang and Willemsen, 2021) or do not use recommendations in their visualization (e.g. Faridani et al., 2010; Biuk-Aghai et al., 2014; Sen et al., 2017; Tintarev et al., 2018). Of these examples that combine both, only few compare InfoVis methods with baseline conditions (e.g. Andjelkovic et al., 2019; Ma et al., 2021). In these few examples that display the entire item space, show recommendations as part of that space, and perform comparative user studies with a baseline system, these baseline systems are realized by masking out the visualization component. We, however, argue that for a fair comparison, a baseline condition should also employ means to browse and explore the item space.

The aim of the study presented here is thus not only to implement different InfoVis techniques in a RS, but also to compare their effects on dependent perceptional variables such as the overview of the item space obtained, the quality of recommendations, their perceived transparency and control, and the user experience in general. As mentioned above, a condition in such a comparative study should be implemented as baseline that allows browsing and exploring of the item space without complex visualizations. Thereby, we deem it especially important to fix components other than the varying visualization between conditions. In particular, these are modalities of interaction, background data, and recommendation engine. In the next chapter, we explain how we constructed our underlying dataset.

4. Hierarchical music dataset

To compare the different item space visualizations we developed a common dataset for which we formulated the following requirements: (1) the dataset should be reasonably large so that it is necessary for users to use interactive tools to explore the item space and RS to cope with information overload; (2) the dataset should consist of *experience products* for which RS are particularly relevant (Senecal and Nantel, 2004; Xiao and Benbasat, 2007); (3) during the experiment, participants should be able to experience the items' media content by listening to it, as item consumption can influence users' perception of recommendations (Loepp et al., 2018); and (4) the dataset should contain meaningful meta-information about the items so that it can be displayed in different ways (e.g. as a two-dimensional map).

Based on these requirements, we developed a dataset in the domain of music with songs as items to be recommended. A music recommender has the advantages that music data is widely available, that songs belong to the category of experience products, which are relatively easy to consume during an online experiment, and that metadata can also be retrieved, e.g. through interfaces of online streaming data sources. Below, we describe the structure of the dataset we used in our experiment and the procedure we employed to construct it. The dataset is organized as a hierarchy because it has been observed that hierarchies of musical genres are perceived as intuitive by users (Lee and Downie, 2004), which may have its origins in physical record stores where records are typically organized into hierarchical levels of genres (Pachet and Cazaly, 2000).

4.1. Data collection

To create a meaningful genre hierarchy, we manually reviewed genres and subgenres listed at *AllMusic*.[5] and *Wikipedia*[6] Based on this, we carefully developed a genre hierarchy with 13 top-level genres and 5 subgenres each (see Table 1). Song, artist, and playlist data were then

[5] https://www.allmusic.com/genres.

[6] https://en.wikipedia.org/wiki/List_of_music_styles.

Table 1
Organization of top-level genres and associated subgenres as used in our dataset. Numbers in parentheses indicate the number of artists that are associated with each genre. Note that there is a certain overlap (i.e. artists can be assigned to multiple genres) and thus the number of artist of the top-level genres is not equal to the sum of artists in its subgenres.

Top-level genre	Subgenres				
blues (510)	*boogie-woogie* (49)	*bountry blues* (110)	*modern blues* (262)	*traditional blues* (150)	*urban blues* (81)
classical (290)	*Baroque* (77)	*Classical Era* (40)	*contemp. classical* (130)	*modernism* (139)	*Romantic Era* (41)
country (846)	*bluegrass* (81)	*contemp. country* (280)	*country pop* (331)	*country rock* (472)	*Texas Country* (123)
electronic (1038)	*dubstep* (127)	*EDM* (413)	*house* (915)	*techno* (810)	*trance* (120)
hip hop (733)	*gangster rap* (200)	*hardcore hip hop* (148)	*southern hip hop* (265)	*trap* (328)	*underground hip hop* (265)
indie (1440)	*Alt Z* (243)	*alternative rock* (616)	*indie folk* (668)	*indiecoustica* (119)	*indietronica* (93)
jazz (533)	*bebop* (144)	*bossa nova* (40)	*contemp. jazz* (163)	*swing* (142)	*vocal jazz* (194)
Latin (653)	*cumbia* (65)	*Latin Pop* (372)	*reggaeton* (149)	*Regional Mexican* (213)	*salsa* (71)
metal (638)	*black metal* (53)	*extreme metal* (318)	*industrial metal* (80)	*nu metal* (176)	*power metal* (77)
pop (989)	*electropop* (341)	*europop* (127)	*neo mellow* (172)	*post-teen pop* (195)	*singer-songwriter* (109)
R&B (949)	*disco* (217)	*funk* (467)	*motown* (193)	*quiet storm* (241)	*soul* (477)
reggae (379)	*dancehall* (164)	*dub* (85)	*lovers rock* (93)	*ska* (114)	*soca* (93)
rock (1434)	*classic rock* (857)	*glam rock* (105)	*psychedelic rock* (135)	*punk rock* (264)	*soft rock* (286)

collected, for which we utilized the Spotify Developer Web API.[7] We made sure that we collected only the data that were strictly necessary to operate our applications. In particular, we performed the following steps to obtain a dataset of artists and songs:

1. We retrieved the 30 categories available at Spotify (e.g. top lists, romance, party).
2. For each of these categories, we fetched up to 100 of the most popular playlists. This resulted in 1802 playlists.[8]
3. We collected all artists that appear at least once in these playlists, resulting in 42 691 different artists.
4. Across all these artists, we gathered the associated genres, resulting in 3959 genres.
5. We aligned our manually created genre hierarchy with the genres obtained from Spotify to match their names (e.g. to match *hip-hop* and *hip hop*). We also added some further genres. For instance, we added all top-level genres to artists associated with a sub-genre (e.g. we added *hip hop* to all artists with the genre *southern hip hop*). A list with all the rules we used for adding these genre associations can be found in the supplement.
6. We removed all artists who were not associated with at least one genre of our hierarchy. This step reduced our dataset to a total of 9063 artists.
7. Finally, we retrieved the 10 most popular songs for each artist. This step left us with 88 948 tracks (sometimes a song appeared in more than one top 10 list). We restricted recommendations to these tracks.

The resulting dataset thus provides a sufficiently large item space that exposes users to information overload and thus creates the need for interactive exploration tools and the personalization capabilities of RS. On the other hand the dataset remains manageable with the computational power at our disposal during data collection, further processing, and the runtime of the experiment.

4.2. Dimensionality reduction

As baseline for the *Map* condition and for determining sample artists for each genre (see Section 4.3), we embedded all artists in a two-dimensional space. To this end, we compared different algorithms for dimensionality reduction (DR) as described in Section 4.2.1. Most of these algorithms need a *dissimilarity matrix* to calculate the two-dimensional embedding. As preparation, we thus computed a dissimilarity score δ_{total} between each pair of artists $a, b \in A$, with A being the set of all artists, as follows:

$$\delta_{total}(a,b) = 0.7 \cdot \delta_{content}(a,b) + 0.2 \cdot \delta_{playlist}(a,b) + 0.1 \cdot \delta_{popularity}(a,b). \quad (1)$$

The three terms in this equation correspond to the three aspects we consider important for a comprehensible dissimilarity between artists: the content of their music (i.e. genres and musical features), their appearance on playlists, and their popularity. The weights of the terms in Eq. (1) and the subsequent equations below (Eqs. (2) and (4)) were determined by a pre-study. For several weight combinations, we calculated a two-dimensional artist embedding as described in Section 4.2. For each of these embeddings, we then asked a small team of researchers to carefully evaluate them in terms of the comprehensibility of the artist distribution—particularly with respect to local artist neighborhoods. Below, we explain how $\delta_{content}$, $\delta_{playlist}$ and $\delta_{popularity}$ were calculated.

Content-based dissimilarity. As the name suggests, $\delta_{content}(a,b)$ aims to capture the content-based dissimilarity between the two artists a and b. The equation we used to calculate this is as follows:

$$\delta_{content}(a,b) = 0.9 \cdot \delta_{genre}(a,b) + 0.1 \cdot \delta_{features}(a,b) \quad (2)$$

with

$$\delta_{genre}(a,b) = 1 - cos(\vec{g}_a, \vec{g}_b) \quad (3)$$

and

$$\delta_{features}(a,b) = 0.7 \cdot \left(\frac{\|\vec{fa}_a - \vec{fa}_b\|}{\max_{x,y \in A}(\|\vec{fa}_x - \vec{fa}_y\|)} \right) + 0.3 \cdot \left(\frac{\|\vec{fs}_a - \vec{fs}_b\|}{\max_{x,y \in A}(\|\vec{fs}_x - \vec{fs}_y\|)} \right). \quad (4)$$

Thereby $\vec{g}_x \in \{0,1\}^{|G|}$ in Eq. (3) denotes a vector with the size equal to that of the set of all genres G (here: $|G| = 3959$). This vector contains a 1 if the corresponding genre is associated with artist x, and a 0 otherwise. In contrast, $\delta_{features}$ represents the dissimilarity of *musical features*. For this, we relied on the six of the 13 available audio features provided by Spotify.[9] More specifically, we chose the following features: *acousticness*, *danceability*, *energy*, *instrumentalness*, *liveness*, and *valence*. We decided to rely on these more abstract attributes because they take into account the other, less abstract features such as *loudness* and *tempo* (e.g. energetic songs are typically fast and loud). Moreover, these six features are all stored in an interval scale of $[0,1]$, which makes them compatible and easy to compare. To calculate a dissimilarity score between artists based on these features, we used Eq. (4). Since Spotify provides such features for each song but not for each artist, we aggregated them, so that $\vec{fa}_x \in \mathbb{R}^f$ denotes a vector containing the average of the features (here: $f = 6$) of the 10 most popular tracks of an artist x. The vector $\vec{fs}_x \in \mathbb{R}^f$, on the other hand, contains the

[7] https://developer.spotify.com/documentation/web-api/.

[8] The number is lower than the possible maximum of 3000 playlists, since 100 playlists were not available for all 30 categories.

[9] For a brief description of all available features, see https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-audio-features.

standard deviation of the features of the 10 most popular tracks of artist x. We decided to use the average as well as the standard deviation of features to consider the typical musical features an artist adheres to but also their musical variation. We normalized the distance between the feature vectors of artists a and b by the maximum distance between any artists' feature vectors.

The weights in Eqs. (2) and (4) were also determined as described above. Here, however, we also based the weighting on theoretical considerations. In this sense, $\delta_{content}$ received the highest weight in Eq. (2), because we wanted the overall distribution to follow the genres of the artists and thus represent the genre hierarchy of our dataset.

Playlist-based dissimilarity. In addition to the content-based dissimilarity explained above, we also considered a dissimilarity based on the co-occurrence of artists in the same playlists. By doing so, we aimed at capturing more subtle similarities between artists that were not necessarily based on the artists' genres or musical features. Additionally, we also wanted to influence the embedding towards the mechanism behind Spotify's recommendation algorithm, in which the co-occurrence of songs in playlists most likely plays a key role (Anderson et al., 2020). We calculate the playlist-based dissimilarity by

$$\delta_{playlist}(a,b) = 1 - \left(\frac{|P_a \cap P_b|}{min(|P_a|,|P_b|)} \right) \tag{5}$$

where P_x denotes the set of all playlists in our dataset that contain at least one song of artist x.

Popularity-based dissimilarity. Finally, we included the popularity of artists in their comparison, but with a rather small weight. This was to capture how "mainstream" artists are, which we also deem an important factor in creating a comprehensible distribution of artists. Each artist is associated with a popularity score by Spotify, which is given in a range of $[0, 100]$ and denoted by $pop(x)$ below. The popularity-based dissimilarity between two artists a and b is calculated as follows:

$$\delta_{popularity}(a,b) = \frac{|pop(a) - pop(b)|}{100}. \tag{6}$$

4.2.1. Comparison of DR algorithms

Different algorithms can be used when reducing the dimensionality of a dataset. To decide for one of the applicable DR algorithms, we compared their scores for *trustworthiness* and *continuity* as described by Kaski et al. (2003): While trustworthiness describes whether close neighborhoods in the projected, i.e. low-dimensional, space can also be found in the original, i.e. high-dimensional, space, continuity describes whether close neighborhoods in the original space can also be found in the projected space. Both scores are in the range of $[0, 1]$, where a perfect value of 1 means that no local neighborhoods are added (trustworthiness) or lost (continuity). Overall, we compared the following algorithms:

- *Multidimensional scaling* (MDS): MDS is a technique that seeks to project a distance matrix onto a low-dimensional embedding. Thereby, MDS tries to preserve the distances between all points in the projection as much as possible to the distances given as input (Borg and Groenen, 2005).
- *Isomap*: Isomap follows a similar technique as MDS, but is not bound to a linear projection of input distances. Instead, Isomap favors maintaining local neighborhoods over global distance settings (Tenenbaum et al., 2000).
- *Locally linear embedding* (LLE): LLE also favors maintaining local distances over global distances. While it, like Isomap, belongs to the class of non-linear DR algorithms, it has no internal global model and maps each local neighborhoods linearly to an output hyperplane (Kayo, 2006).
- *T-distributed stochastic neighbor embedding* (t-SNE): In general, t-SNE also favors maintaining local relationships between data points, but it has been shown to be superior at revealing clusters in the data while also preserving global relationships to a certain degree. It is thus well suited for both global and local representations of high-dimensional datasets in a low-dimensional space (van der Maaten and Hinton, 2008).

Table 2
Scores for trustworthiness and continuity for the DR algorithms we used.

	t-SNE	*MDS*	*LLE*	*Isomap*
Trustworthiness	0.998	0.910	0.885	0.921
Continuity	0.993	0.824	0.950	0.947

Results on trustworthiness and continuity for these algorithms can be found in Table 2.

We noted that t-SNE yielded the highest scores for trustworthiness and continuity, and therefore concluded that this method was best able to preserve the structure of the dataset. In addition, we judged the two-dimensional artist distribution as obtained from t-SNE to be comprehensible, and thus decided to stick to this distribution for all subsequent steps.

4.3. Determination of representative samples

For each genre, we determined two representative sample artists to illustrate the corresponding genre. More specifically, for each genre $i \in G$, we determined two artists, one at the center of the cluster of all artists associated with that particular genre by

$$rep\_center(i) = \arg\max_{x \in A_i} \left(\frac{pop(x)}{100} + \left(1 - \frac{\|\vec{p}_x - \vec{c}_i\|}{\max_{y \in A_i}(\|\vec{p}_y - \vec{c}_i\|)} \right) \right), \tag{7}$$

and one that lies in its periphery by

$$rep\_periphery(i) = \arg\max_{x \in A_i} \left(\frac{pop(x)}{100} + \frac{\|\vec{p}_x - \vec{c}_i\|}{\max_{y \in A_i}(\|\vec{p}_y - \vec{c}_i\|)} \right). \tag{8}$$

Here $\vec{p}_x \in \mathbb{R}^2$ denotes the two-dimensional position vector of artists x as calculated by t-SNE as explained above. The set A_i represents the set of all artists associated with genre i. The vector $\vec{c}_g \in \mathbb{R}^2$ denotes the *centroid* of genre g, i.e. the arithmetic mean of all artist positions associated with that genre. Finally, $pop(x)$ corresponds to the popularity score of artist x, as introduced in Section 4.2 and Eq. (6). With these two samples for each genre, we wanted to present one very typical artist for a genre, but also one that allows users to perceive its variety. This is especially important for very broad genres such as *Latin*, which includes typical Latin music such as *salsa* on the one hand, but also more exotic subgenres such as *reggaeton*. All samples determined by this method can be found in the supplement.

5. Prototypes for interactive item space exploration

In the user study described in Section 6, we compared three prototypes that depict the item space of music artists and songs introduced in the previous chapter. Each of these prototypes is implemented as a web application and uses a unique style for visualizing the item space: A very simple *List* visualization, which represents the item space as a hierarchical tree list and serves us as baseline condition. A *Treemap* visualization, where genres and subgenres are displayed as tiles. And a *Map* visualization that uses the metaphor of a geographical map, representing artists as discrete points on a two-dimensional surface.

Except for the different visualization of the item space, all our prototypes follow the same general structure (see Fig. 1). The visualization is displayed in a large central exploration area (A). Below, details of the currently selected item (i.e. artist) can be displayed on demand in the details area (B). In addition, all prototypes share

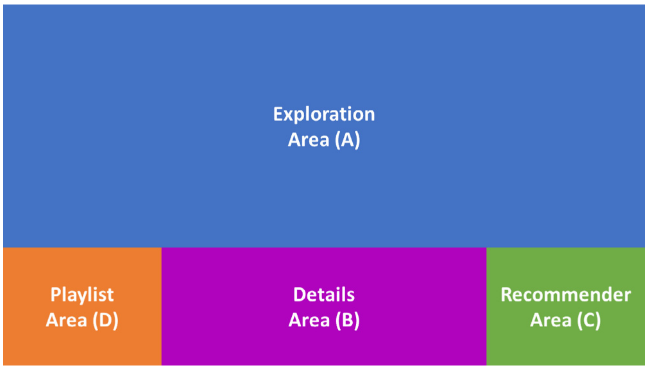


Fig. 1. The main structure that our three prototypes have in common: The exploration area (A) contains the visualization of the item space, which varies among prototypes. In each prototype it also contains a search bar and a help button in the upper right corner. The details area (B) shows the currently selected artist along with their top 10 songs and controls for play/pause, like, and add a song to the playlist. As soon as the first song is liked, the recommendation area (C) depicts a list of 10 song recommendations. Finally, the playlist area (D) can be used to create and manage a custom playlist.

areas for personalized recommendations (C) and creating a playlist (D). These three components are presented in Fig. 2. To further blend search and browsing capabilities, each exploration area also contains a search bar in which users can type in a name of an artist, who is then highlighted in the exploration area and displayed in the details area. The exploration area also contains a help button that opens a pop-up window showing instructions how to use the application. In the details area, there is a button to express preferences for songs which will result in an immediate recalculation of recommendations. During the design of each prototype, we followed the *information seeking mantra* by Shneiderman (1996).

All prototypes are implemented as single-page web application. To determine distinctive colors of the genres, we utilized *colorbrewer*.[10] Recommendations are calculated by Spotify: each time a user clicks the like-button, a request for recommendations is sent to the Spotify API with all previously liked songs as seed items. Further details on the interface design of each prototype can be found below while information on how we implemented them can be found in the supplementary material.

5.1. List prototype

Depicting hierarchies in the form of a (tree) list constitutes a common method. For this reason, we decided to use an alphabetically ordered, vertically displayed list as our baseline condition (Fig. 3). In particular, for each top-level genre, one row is displayed in this list. These rows can be expanded by clicking on them, whereupon the corresponding subgenres are displayed. When these subgenres are selected, they also expand, and reveal all artists associated with that subgenre. Note that artists may be assigned to multiple genres and hence appear in multiple subgenre lists. When users search for a specific artist using the search bar in the upper right corner, that artist's first matching subgenre and associated supergenre expand. Then, the viewport scrolls to center that subgenre.

5.2. Treemap prototype

In a treemap, the space available for visualization is represented as rectangles for each element at each level of a hierarchical data structure to be visualized. The area of the rectangles thereby indicates a quantifiable attribute of the corresponding element. We apply this method to display genres and subgenres as elements and the number of artists they contain as quantifiable attribute (Fig. 4). More specifically, we follow the approach of squarified treemaps (Bruls et al., 2000), where rectangles are arranged so that their aspect ratios are as balanced as possible. Instead of thin, elongated rectangles the algorithm thus aims to create squares, which are easier to interpret. In our visualization, the resulting tiles (i.e. squarified rectangles) are sorted in decreasing size from top left to bottom right. In other words, the tiles start with the most popular genre (i.e. the genre with the most associated artists) at the top left and end with the least popular genre at the bottom right. When users click on one of these tiles, the view "zooms in" so that the display is now filled with the subgenres of the genre clicked on. The same rules apply to these subgenres in terms of their size and position as to the genres. Within the tile of each sugrenres, an alphabetically ordered, scrollable list of all artists related to that subgenre is depicted.

5.3. Map prototype

The prototype with the most complex visualization depicts the item space as a two-dimensional map (Fig. 5). All artists are embedded in this two-dimensional space as described in Section 4.2. Afterwards, to create a more uniform distribution (i.e. a high entropy layout), we applied a small magnetic force based on *Rutherford-scattering* to the artists. As a result, they repel each other slightly so that very densely clustered areas become more readable.

The three levels (top-level genres, subgenres, all artists) of the genre hierarchy were implemented in the *Map* application with a smooth transition. In this way, our application applies *semantic zooming*: the further the user zooms in, the more items and their labels become visible in the respective area. In addition to zooming in and out, users can pan the map horizontally and vertically to explore neighboring regions. To search for a particular artist, users can utilize the search bar in the upper right corner. The artist found is selected immediately (i.e. shown in the details view) and the map is centered on that artist, keeping the current zoom level constant. The minimap in the upper left corner indicates at any time the current position of the view on the main map. By clicking on this minimap, users can have the main view centered at this position.

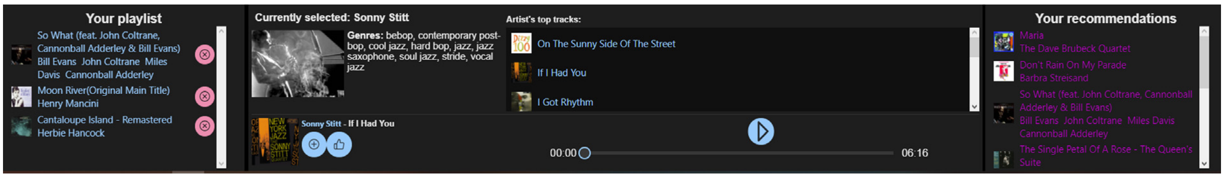
5.4. Visual complexity of prototypes

Comparing complex visual interfaces is not trivial. In an attempt to quantify the complexity of our user interfaces nonetheless, we manually counted all artists visible in screenshots we took of each condition (see supplementary material) and measured their *visual clutter*. To achieve the latter, we calculated the *feature congestion* (FC) value, which measures how "difficult it would be to add a new item that would reliably draw attention", and the *subband entropy* (SE), which represents "the number of bits required for [. . .] image coding" (Rosenholtz et al., 2007).

With 39 visible artists, the *List* interface displays the lowest number of items simultaneously. Accordingly, the clutter measures are also lower than for the other interfaces (SE = 1.59, FC = 2.06). There are 95 artists visible on the screenshot of the *Treemap* interface. The corresponding clutter measures are: SE = 2.17, FC = 2.78. Finally, the screenshot of the *Map* interface shows the highest number of artists (305). While the SE is also highest for this interface (2.77), the corresponding FC is slightly lower than for the *Treemap* condition (2.26).

While we acknowledge that these values are based on manually taken screenshots and are therefore rather illustrative, we have carefully attempted to apply the same hierarchical level of detail to all of them and thus deem the results reported here to be reasonably valid. However, future research should definitely investigate the comparison of clutter between different visualizations of the item space.

[10] https://colorbrewer2.org.



Your playlist

So What (feat. John Coltrane, Cannonball Adderley & Bill Evans)
Bill Evans John Coltrane Miles Davis Cannonball Adderley

Moon River(Original Main Title)
Henry Mancini

Cantaloupe Island - Remastered
Herbie Hancock

Currently selected: Sonny Stitt

Genres: bebop, contemporary post-bop, cool jazz, hard bop, jazz, jazz saxophone, soul jazz, stride, vocal jazz

Artist's top tracks:

On The Sunny Side Of The Street
If I Had You
I Got Rhythm

Sonny Stitt - If I Had You

00:00 06:16

Your recommendations

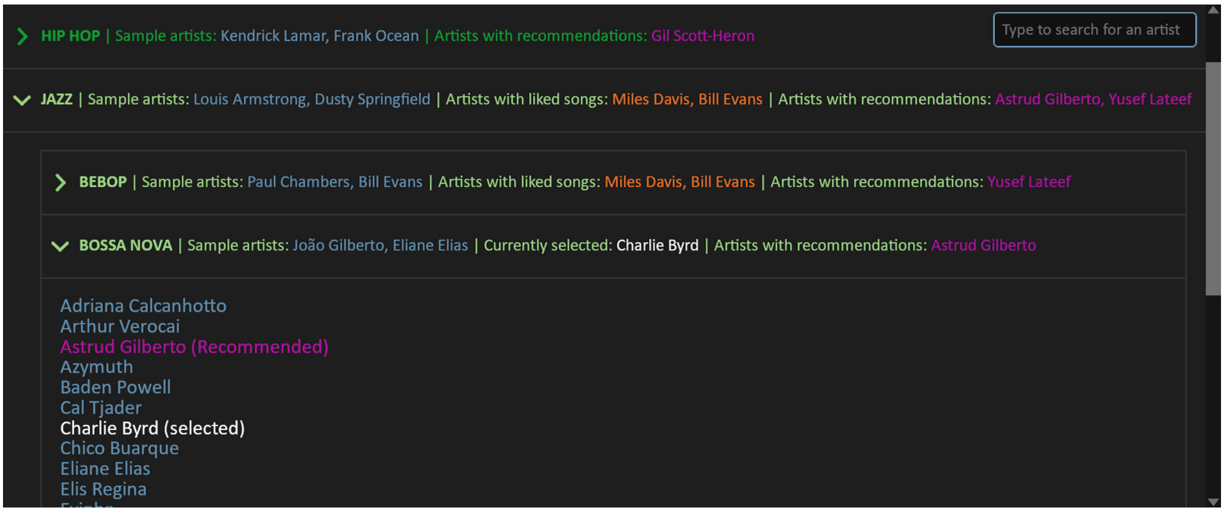
Maria
The Dave Brubeck Quartet

Don't Rain On My Parade
Barbra Streisand

So What (feat. John Coltrane, Cannonball Adderley & Bill Evans)
Bill Evans John Coltrane Miles Davis Cannonball Adderley

The Single Petal Of A Rose - The Queen's Suite

Fig. 2. Screenshot of the control elements as they were used in all three prototypes. The central area shows the currently selected artist and the currently selected song. It also contains controls of the music player. The area on the right hand side shows a list with current recommendations while the area on the left lets users manage a custom playlist.



HIP HOP | Sample artists: Kendrick Lamar, Frank Ocean | Artists with recommendations: Gil Scott-Heron

Type to search for an artist

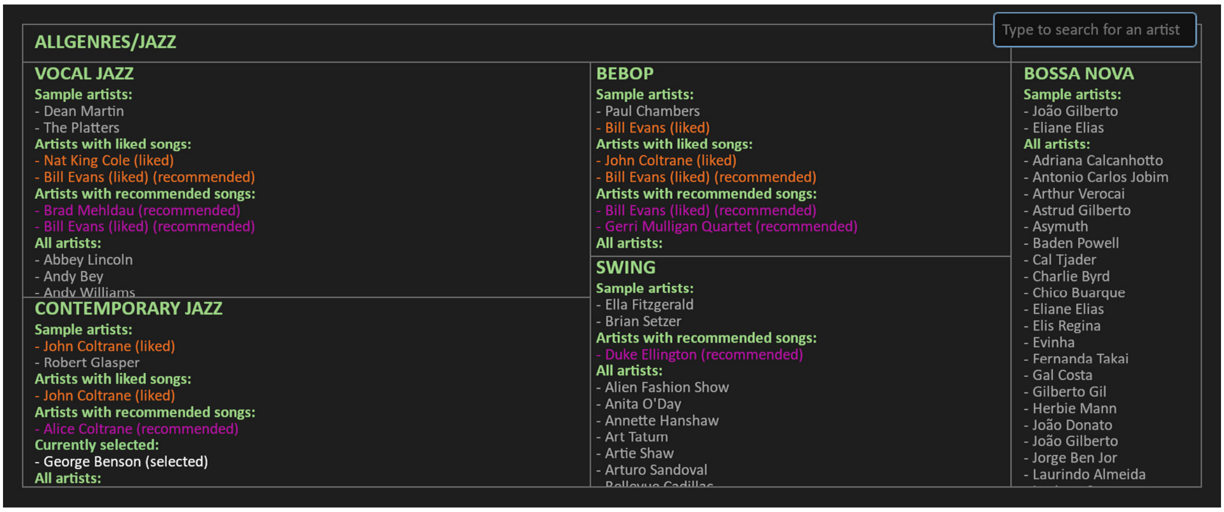
JAZZ | Sample artists: Louis Armstrong, Dusty Springfield | Artists with liked songs: Miles Davis, Bill Evans | Artists with recommendations: Astrud Gilberto, Yusef Lateef

BEBOP | Sample artists: Paul Chambers, Bill Evans | Artists with liked songs: Miles Davis, Bill Evans | Artists with recommendations: Yusef Lateef

BOSSA NOVA | Sample artists: João Gilberto, Eliane Elias | Currently selected: Charlie Byrd | Artists with recommendations: Astrud Gilberto

Adriana Calcanhotto
Arthur Verocai
Astrud Gilberto (Recommended)
Azymuth
Baden Powell
Cal Tjader
Charlie Byrd (selected)
Chico Buarque
Eliane Elias
Elis Regina
Evinha

Fig. 3. Print-optimized screenshot of the exploration area in the *List* prototype: The item space is depicted as a hierarchical list. The top-level genre *jazz* is currently expanded and its subgenres are visible. Of these, the subgenre *bossa nova* is expanded. The artist *Charlie Byrd* is currently selected (white font) and artists with liked songs (orange font), and recommendations (magenta font) are showing.



ALLGENRES/JAZZ

Type to search for an artist

VOCAL JAZZ
Sample artists:
- Dean Martin
- The Platters
Artists with liked songs:
- Nat King Cole (liked)
- Bill Evans (liked) (recommended)
Artists with recommended songs:
- Brad Mehldau (recommended)
- Bill Evans (liked) (recommended)
All artists:
- Abbey Lincoln
- Andy Bey
- Andy Williams

CONTEMPORARY JAZZ
Sample artists:
- John Coltrane (liked)
- Robert Glasper
Artists with liked songs:
- John Coltrane (liked)
Artists with recommended songs:
- Alice Coltrane (recommended)
Currently selected:
- George Benson (selected)
All artists:

BEBOP
Sample artists:
- Paul Chambers
- Bill Evans (liked)
Artists with liked songs:
- John Coltrane (liked)
- Bill Evans (liked) (recommended)
Artists with recommended songs:
- Bill Evans (liked) (recommended)
- Gerri Mulligan Quartet (recommended)
All artists:

SWING
Sample artists:
- Ella Fitzgerald
- Brian Setzer
Artists with recommended songs:
- Duke Ellington (recommended)
All artists:
- Alien Fashion Show
- Anita O'Day
- Annette Hanshaw
- Art Tatum
- Artie Shaw
- Arturo Sandoval
- Bellevue Cadillac

BOSSA NOVA
Sample artists:
- João Gilberto
- Eliane Elias
All artists:
- Adriana Calcanhotto
- Antonio Carlos Jobim
- Arthur Verocai
- Astrud Gilberto
- Asymuth
- Baden Powell
- Cal Tjader
- Charlie Byrd
- Chico Buarque
- Eliane Elias
- Elis Regina
- Evinha
- Fernanda Takai
- Gal Costa
- Gilberto Gil
- Herbie Mann
- João Donato
- João Gilberto
- Jorge Ben Jor
- Laurindo Almeida

Fig. 4. Print-optimized screenshot of the exploration area in the *Treemap* prototype: The item space is depicted as a squarified Treemap. The top-level genre *jazz* is currently open, so that its subgenres are visible. The artist *George Benson* (white font) is currently selected. In addition, artists with liked (orange font) and recommended (magenta font) songs are being displayed for the currently visible genres.

6. User study

To compare the different visualization methods, we conducted an empirical user study using a between-subjects design. Consistent with our RQs, we aimed at investigating the influence of the item space visualization on the perception of recommendations, differences in how participants explored the item space, and connections between user characteristics and these dimensions. The study was performed as an online survey that included interaction tasks with the prototypes and different questionnaires. As questionnaire tool, we used the online platform *SoSci Survey*.[11] The study was approved by the local ethics committee of the *University of Duisburg–Essen* in Germany.

6.1. Preparation and survey of user characteristics

Participants were briefed that they were about to test a novel application: the *MusicExplorationApp*. They were then asked for their consent

[11] https://www.soscisurvey.de/.

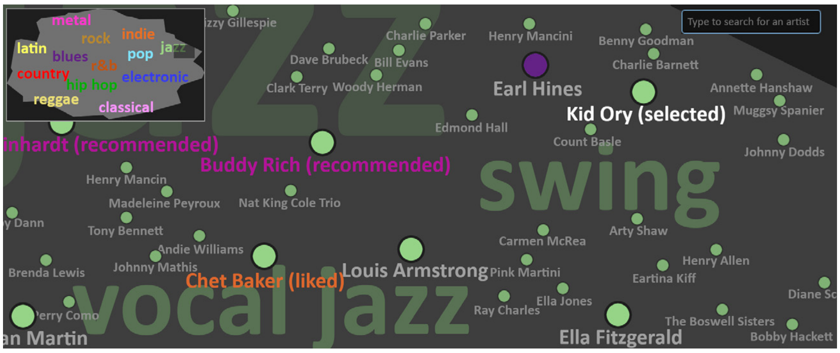
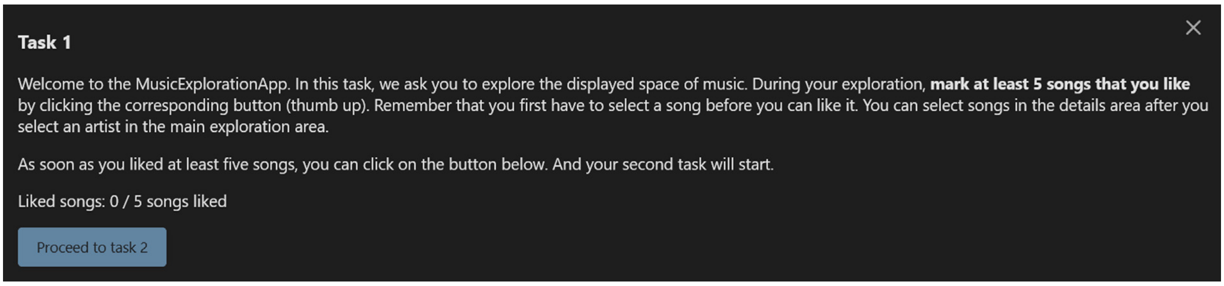


Fig. 5. Print-optimized screenshot of the exploration area in the *Map* prototype: The item space is depicted as topological map. The view is currently on an intermediate zoom level and focuses on the subgenres *swing* and *vocal jazz* of the top-level genre *jazz*. The artist *Kid Ory* (white font) is currently selected. Artists with liked (orange font) or recommended (magenta font) songs are highlighted on their respective map position. The minimap in the upper left corner displays the position of the current view (transparent black rectangle on the east) in context of the entire map.



Task 1

Welcome to the MusicExplorationApp. In this task, we ask you to explore the displayed space of music. During your exploration, **mark at least 5 songs that you like** by clicking the corresponding button (thumb up). Remember that you first have to select a song before you can like it. You can select songs in the details area after you select an artist in the main exploration area.

As soon as you liked at least five songs, you can click on the button below. And your second task will start.

Liked songs: 0 / 5 songs liked

Fig. 6. Screenshot of the description of task 1 as it was presented to participants.

to use their data in anonymized form and redirected to an initial questionnaire. This questionnaire was composed of questions about participants' demographics, *Big Five* personality traits, which we measured with the 10-item inventory by Rammstedt and John (2007), and *decision style*, which was measured with the instrument by Hamilton et al. (2016). This was followed by an online implementation[12] of the *Corsi's block-tapping test* (Corsi, 1972) to measure *visual memory* capacity. Subsequently, participants were randomly assigned to one of the three conditions (i.e. *List*, *Treemap*, *Map*) and were given a brief introduction to that system in textual form, including several screenshots (included in the supplement). It was made clear to them which interaction options they have, that recommendations are calculated each time they like a song, and that Spotify's recommendation algorithm is used. Participants were then directed to the respective system of the condition to which they were assigned. Here, they first had to login with their Spotify credentials before they were presented with the application. For the purpose of this study, each prototype contained an additional area at the top of the screen where users could read the description of the current task. This description was collapsible into a thin bar at the top of the screen. This bar permanently showed the title of the current task, the progress made, and a button to proceed to the next task. Below we describe each task and our motivation behind it. We also indicate below which areas and interactive elements were disabled and enabled for each task.

6.2. Task 1

The first task served as introduction to the system and to elicit an initial preference profile of the active participant (Fig. 6). It also aimed at inducing them to actively explore the item space. In this task, participants had to rate 5 songs. On minimizing the long task description in Fig. 6, the task title ("*Like 5 songs*") and the counter remained visible. The playlist area, add-to-playlist button, and recommendation area were disabled during this task. All other elements and areas were enabled. Liking a song did not trigger an immediate calculation of recommendations. After a participant had liked at least five songs, they were able to proceed to the next task.

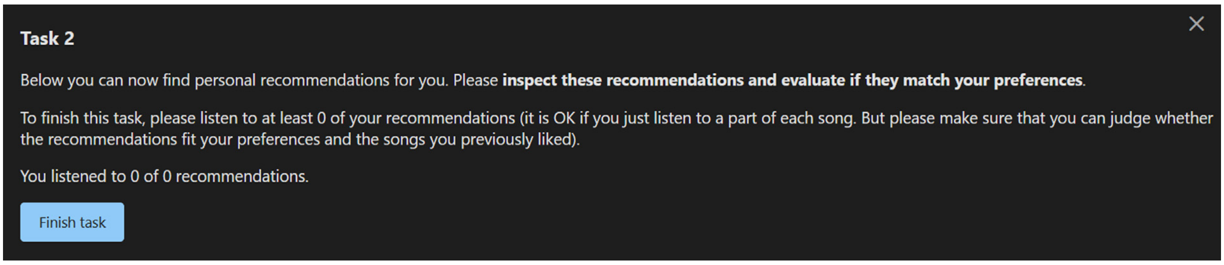
6.3. Task 2

In this task, participants were asked to listen to their recommendations (Fig. 7). With this task, we wanted to draw participants' attention to the recommendation part of the system. In contrast to task 1, the recommendation area was enabled during this task. It displayed a list of recommendations based on the preferences expressed in task 1. In addition, the recommendations were also shown inside the exploration area, as explained for each prototype in Section 5. The playlist area, the add-to-playlist button, and the like button were disabled during this task. Thus, participants were not able to trigger a recalculation of recommendations. After selecting each of their 10 recommendations at least once, participants could proceed to the next task.

6.4. Task 3

While task 1 and task 2 served as an introduction and aimed to familiarize participants with the system, its exploration area, and recommendations, tasks 3 and 4 aimed at simulating a more realistic day-to-day situation in which a user creates a new playlist. To this end, in task 3, we asked participants to create a playlist of at least six songs for themselves (Fig. 8). From task 3, all functionalities of the system and all areas of the interface were available. The same recommendations as in task 2 were shown. As soon as the playlist contained at least six songs, participants were able to proceed to the next task.

[12] We used a Javascript adaption of Professor Gijsbert Stoet's implementation (https://www.psytoolkit.org/experiment-library/corsi.html).



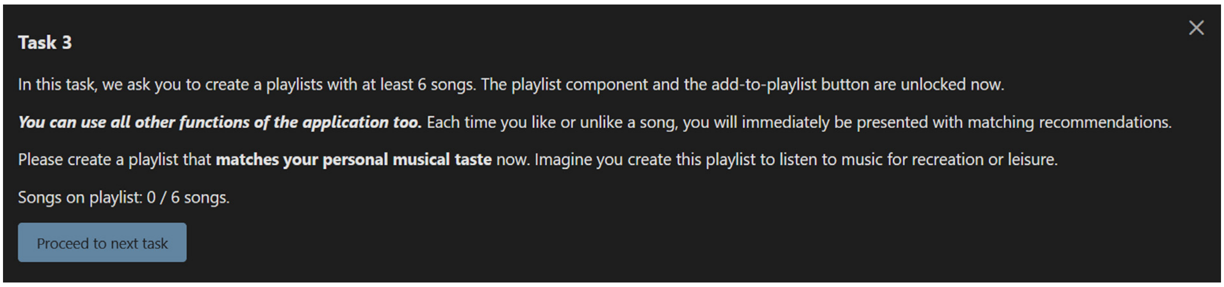
Task 2

Below you can now find personal recommendations for you. Please **inspect these recommendations and evaluate if they match your preferences**.

To finish this task, please listen to at least 0 of your recommendations (it is OK if you just listen to a part of each song. But please make sure that you can judge whether the recommendations fit your preferences and the songs you previously liked).

You listened to 0 of 0 recommendations.

Fig. 7. Screenshot of the description of task 2 as it was presented to participants.



Task 3

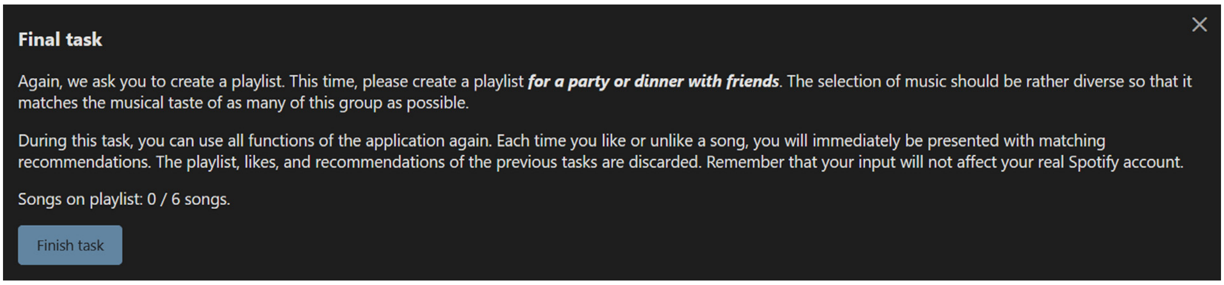
In this task, we ask you to create a playlists with at least 6 songs. The playlist component and the add-to-playlist button are unlocked now.

You can use all other functions of the application too. Each time you like or unlike a song, you will immediately be presented with matching recommendations.

Please create a playlist that **matches your personal musical taste** now. Imagine you create this playlist to listen to music for recreation or leisure.

Songs on playlist: 0 / 6 songs.

Fig. 8. Screenshot of the description of task 3 as it was presented to participants.



Final task

Again, we ask you to create a playlist. This time, please create a playlist ***for a party or dinner with friends***. The selection of music should be rather diverse so that it matches the musical taste of as many of this group as possible.

During this task, you can use all functions of the application again. Each time you like or unlike a song, you will immediately be presented with matching recommendations. The playlist, likes, and recommendations of the previous tasks are discarded. Remember that your input will not affect your real Spotify account.

Songs on playlist: 0 / 6 songs.

Fig. 9. Screenshot of the description of task 4 as it was presented to participants.

6.5. *Task 4*

As in the previous task, in task 4, the participants were asked to create a playlist (Fig. 9). This time, however, they were to create a playlist for a party or dinner with friends. With task 4, we aimed at letting users think about the diversity of their playlist and thus make more use of the exploratory components of the application. As in task 3, all functionalities of the system and all areas of the interface were available during task 4. After this playlist also consisted of at least six songs, participants were directed to a final questionnaire.

6.6. *Final questionnaire*

The final questionnaire contained questions on various dimensions regarding the perception of recommendations, which we used to test how the visualization of the item space affects how users perceive their recommendations. In particular, we used questions to survey *recommendation quality* and *recommendation variance* (5-point Likert scale) from the instrument introduced by Knijnenburg et al. (2012), while perceived *recommendation transparency* and *recommendation novelty* were measured by items from the *ResQue* inventory (5-point Likert scale, Pu et al. (2011)). Furthermore, we added the *user experience questionnaire* (UEQ), for which we used the shortened UEQ version presented by Schrepp et al. (2017) with a 7-point bipolar scale. The UEQ comprises *pragmatic* and *hedonic* UX, which enables us to make statements about experienced system usability as well as other user experiences such as excitement and interest. We complemented these existing questionnaires with some self-generated questions. Specifically, these were questions assessing the participants' perceived *degree of overview* over the item space (5 items, Cronbach's α = .746) and perceived *degree of control* over the recommendation process (3 items, Cronbach's α = .738), both of which were measured on a 5-point Likert scale. The exact wording of all instruments used can be found in the supplement of this article.

6.7. *Sample*

We recruited 91 (51 male, 38 female, 2 other) participants with an age of M =26.53 (SD = 7.07) through the crowdworking platform *Prolific*.[13] As mandatory technical requirements, we specified that participants needed a Spotify Premium account, a desktop computer with a mouse and speakers or headphones, and a modern browser with Javascript enabled. We paid participants £ 5 (about $ 6.80 or € 5.96). To complete the study, participants needed an average of 33.17 min (SD = 18.32). Since Prolific is a UK-based platform, it was not surprising that most participants indicated that they currently liven in the United Kingdom (20.9 %), followed by other European countries such as Portugal (12.1 %) and Poland (11.0 %). However, participants from outside Europe took also part in the study. There were, for instance, 6.6 % participants from Canada and 5.5 % from Mexico. The sample was rather highly educated, with most participants reporting a university

[13] https://www.prolific.co/.

Table 3
Descriptive questionnaire results for perceptional variables of the three prototypes. In case of a significant difference, the higher value is marked in bold and the condition to which the significant difference was present is indicated by a superscript letter. Pragmatic and hedonic user experience differed on a significance level <.05, and recommendation novelty on a significance level <.1. Exact p-values can be found in Section 7.1.

	List		*Treemap*		*Map*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Perception of application/item space visualization						
Pragmatic user experience	**5.28**^{T}	1.29	4.66	1.48	5.02	1.41
Hedonic user experience	4.65	1.15	4.38	1.32	**5.40**^{L,T}	1.09
Item space overview	3.55	0.70	3.30	0.81	3.61	0.77
Perception of recommendations						
Quality	4.07	0.64	3.96	0.73	4.11	0.69
Variety	3.35	0.90	3.65	0.66	3.60	0.71
Novelty	3.10	0.16	3.47	0.87	**3.60**^{L}	0.94
Transparency	4.07	0.86	4.00	0.79	4.25	0.82
Degree of control	3.66	0.83	3.87	0.79	3.87	0.81

degree (39.6 %) as their highest level of education, followed by a high school diploma (24.2 %), and a higher education entrance qualification (15.4 %). As their profession, 40.7 % reported to be employed, while 29.7 % responded being a university student, and 12.1 % responded being unemployed or seeking employment.

Of the 33.17 min, participants took for completing the entire study, they spent an average of 15.54 min (SD = 08.68) for interacting with the application and 17.73 min (SD = 14.62) for processing the questionnaires.

7. Results

This chapter is structured according to our three research questions introduced in Section 1: Regarding RQ1 (*How does the type of item space visualization influence users' perception and understanding of the item domain and the recommendations?*), we present a quantitative analysis of questionnaire responses (Section 7.1), followed by a mediation analysis in Section 7.2, which we used to unravel causal relationships between *hedonic user experience* and *recommendation novelty*. We furthermore present a brief report on qualitative comments in Section 7.3. In terms of RQ2 (*How does the type of item space visualization influence how users interact with this visualization and the recommendations?*), we present log data in Section 7.4. These data are further analyzed in Section 7.5, where we explore interaction patterns and present an alternative perspective on our results by clustering participants based on how they interacted with the application. Finally, in Section 7.6, we analyze user characteristics according to RQ3 (*How do users' characteristics influence how they perceive, understand, and interact with the item space and the recommendations?*).

Throughout this chapter, we use an α-level of 0.1 for all significance tests. We consider this threshold acceptable since we are comparing relatively complex visualizations which is, despite the experimental rigor applied, to some extent still exploratory. However, we also report exact p-values for each of these tests.

7.1. Questionnaire results

Descriptive results for the perceptual dimensions of item space visualization and recommendations as introduced in Section 6.6 can be found in Table 3. To determine whether there was a difference between our three conditions in terms of the results in Table 3, we conducted a one-way MANOVA. The multivariate effect for *condition* was significant ($F(16, 162) = 2.558$, $p = .002$, *Wilk's* $\Lambda = .637$, $\eta_p^2 = .202$). Post hoc tests with Bonferroni-corrected α-levels revealed significant differences regarding the *recommendation novelty* ($F(2, 88) = 4.025$, $p = .074$, $\eta_p^2 = .058$) between the *Map* and the *List* condition ($p = .082$). Furthermore, we found significant differences in terms of *pragmatic user experience* ($F(2, 88) = 7.384$, $p = .026$, $\eta_p^2 = .079$) between the *List* and the *Treemap* condition ($p = .028$), and in terms of *hedonic user experience* ($F(2, 88) = 8.524$, $p = .003$, $\eta_p^2 = .121$), between the *List* and *Map* condition ($p = .049$) and between the *Treemap* and *Map* condition ($p = .003$).

Apart from this, we found no differences with statistical significance in our questionnaire results. However, we did find minor trends in some descriptive results. For instance, we observed that the perceived *degree of control* was slightly higher in the *Map* and *Treemap* condition compared to the *List* condition. *Degree of overview* and *recommendation transparency* tended to be higher in the *Map* condition compared to the *Treemap* condition (and to a lesser extent in the *List* condition).

7.2. Mediation analysis

As reported above, we found an unexpected difference in the perceived *recommendation novelty* between conditions. We hypothesize that the *hedonic user experience* is the reason for this perceived novelty of recommendations. We base this hypothesis on three indicators:

1. *Recommendation novelty* was highest in the condition with the *highest hedonic user experience* (i.e. the *Map* condition). However, recommendations in each condition were provided by the Spotify API and thus calculated in the exact same way.
2. *Recommendation novelty* as subjectively perceived by participants does not correlate with an objective novelty of recommendations as calculated by the *self-information* (Eq. (9)).
3. Prior work has shown that users judge the visual appeal of websites in the first moments when accessing it Lindgaard et al. (2006). We thus suppose that participants assessed the *hedonic user experience* during task 1 and thus *before* they were presented with recommendations. We argue further that participants cognitively inferred the novelty of recommendations based on this judgment, as has been observed in situations with incomplete information (Kardes et al., 2004).

To test this hypothesis we conducted a mediation analysis. More specifically, we used a bootstrapping approach (Hayes, 2022) to test the hypothesized mediation model depicted in Fig. 10. First, we created a binary dummy variable ($condition_{\text{map}}$) for the *Map* condition (1 = *Map*; 0 = *List* or *Treemap*). Subsequently, we utilized the "PROCESS" macro (Hayes, 2022) in SPSS with 90 % confidence intervals and bootstrapping with $N = 5000$ to test for significance of an indirect effect mediated by *hedonic user experience*. We found the path $condition_{\text{map}}$ → *hedonic user experience* to be significant ($t(89) = 3.37$, $p = .001$). Also, the path *hedonic user experience* → *recommendation novelty* was significant ($t(89) = 5.90$, $p < .001$). Through this mediation, the direct effect of $condition_{\text{map}}$ → *recommendation novelty* disappears ($t(89) = -0.19$, $p = .851$), revealing that the relation of $condition_{\text{map}}$ and *recommendation novelty* is fully mediated by *hedonic user experience*.

7.3. Qualitative results

To analyze qualitative results, we studied comments participants made at the end of the questionnaire. Therefore, we first ordered participants' comments according to the condition they were in. Subsequently, two researchers checked those comments and discussed them regarding their core statement. Below we briefly report our findings. All comments in their original wording can be found in the supplement.

Participants who were exposed to the *List* application gave a lot of positive feedback. "*It is pretty good*" (P46), "*I genuinely think this app is great!*" (P35), and "*I liked the app!*" (P48) are examples. Others, however, did not like this condition as much, commenting statements such as "*Design could be better*" (P91), "*it looks not beginner user friendly.*" (P65), or "*The system is a little slow*" (P88). Many participants compared the app to their experience with Spotify: "*I was surprised how easy*

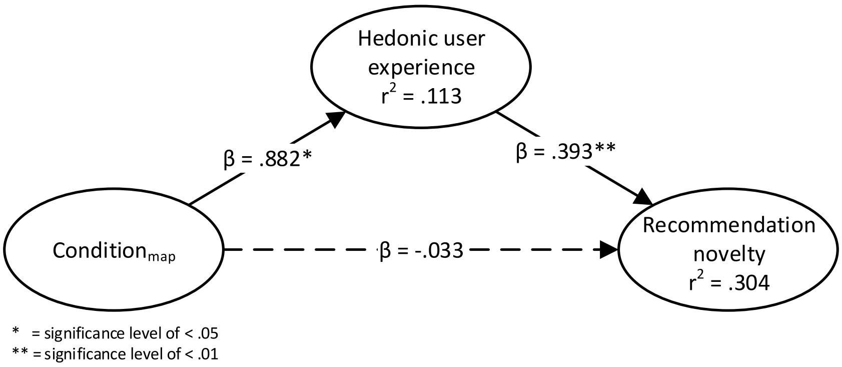


Fig. 10. Pathway diagram of the mediation of *hedonic user experience* in the relationship between *condition*$_{\text{map}}$ and perceived *recommendation novelty*. The variable *condition*$_{\text{map}}$ is a dummy variable coded as binary variable with respect to the *Map* condition (1 = *Map*; 0 = *List* or *Treemap*).

it was to use compared to spotify" (P53), "*[the system] was capable of recommending new songs for me [...], slightly better than the 'spotify radio' or 'people like you also played'*" (P71).

In the *Treemap* condition, we found one participant who made a similar statement: "*I added a few songs from this to my spotify playlist! On spotify my recommendations are kinda boring and not what I really like.*" (P57). In this condition, the recommendations were also mainly experienced as positive. However, the general perception of the interface in the *Treemap* condition was not that positive: "*I liked the recommendation part but the app in general was confusing*" (P57). Some participants pointed at general problems with usability, such as P51: "*the space to see the songs is too small.*" Some also complained about the performance in computing recommendations and described the system as "*slow*" (P26), or taking "*a bit of time to load*" (P18). That being said, many participants enjoyed using the system and commented statements such as "*the experience was pretty nice*" (P43), "*[the system] is fun to use! No need to improve anything.*" (P64), or "*this system was well put together! It worked perfectly and did its job.*" (P41).

The most positive statements were made by participants in the *Map* condition. They gave comments like "*It seems very interesting! I'd love to use a tool like this regularly*" (P74), "*I found the application very intuitive*" (P28), "*It is amazing! This is exactly what i was looking for*" (P39), and "*It was so cool!! I really enjoy it*" (P23). On the other hand, the *Map* interface was also sometimes described as "*very laggy*" (P10), "*a bit confusing*" (P03), or just "*poorly made*" (P68). Some stated that they "*did not feel like exploring the genre map at all*" (P62), and thus wished that "*the 3 boxes at the bottom [i.e. recommendation, details, and playlist area] were taller*". Finally, we received some conceptual suggestions to consider in the next iteration of development: "*I'd be interested to see [my taste profile] as a "lit up" visual map just for me and [to] be able to compare with others.*" (P70), "*It needs to look more modern in my opinion. [...]Maybe have the world map look more like the world?*" (P47).

7.4. Log data

While participants interacted with the different prototypes, we logged various events, some of which can be found in Table 4.[14] To test for differences in interaction behavior and thus to answer our second research question, we conducted another one-way MANOVA and post hoc tests with Bonferroni correction. The multivariate effect for *condition* was significant ($F(18, 160) = 2.587$, $p = .001$, *Wilk's* $\Lambda = .600$, $\eta_p^2 = .225$). Participants took a comparable amount of time to complete their tasks across conditions (no significant differences). Only in tendency, participants in the *Treemap* condition seem to take slightly less time to complete their tasks. We also analyzed in which region of the interface each click occurred: exploration area, details area, recommendation area, and playlist area (see also Fig. 1). Descriptive data can be found in Table 4. In summary, we only found that the *number of clicks in the recommendation area* was significantly higher in the *Map* than in the *Treemap* condition ($F(2, 88) = 2.788$, $p = .074$, $\eta_p^2 = .060$). However, when counting interactions for all areas and considering mouse scrolls too, this *number of overall interactions* was significantly higher in the *Treemap* condition than in the *List* condition ($F(2, 88) = 6.627$, $p = .002$, $\eta_p^2 = .131$). Apparently, participants in the *Treemap* condition performed many scroll interactions, while they used less recommendations for completing their tasks. The latter is further backed up by the observation, that participants differed in terms of their *ratio of recommendations on playlist* during task 3 and 4 ($F(2, 88) = 4.126$, $p = .019$, $\eta_p^2 = .086$) between the *Map* and *Treemap* condition ($p = 0.072$), and the *Treemap* and *List* condition ($p = .029$). Apart from this, there were no differences with statistical significance, only some minor tendencies as depicted in Table 4.

To make objective statements about the novelty of recommendations, we calculated the *self-information* (Zhou et al., 2010; Vargas and Castells, 2011) of the set of recommended items, which is defined as the averaged negative logarithm of each item's popularity:

$$Novelty(R) = \frac{\sum_{i \in R} -\log_2(\frac{pop(i)}{100})}{|R|} \qquad (9)$$

The results for the novelty of the recommendations (this is a mean value over each recommendation list received by a participant) can be found in the last row of Table 4. The values are very similar over all conditions, and no significant difference can be found. Thus, we conclude that the *perceived* novelty of recommendations, as elicited by the questionnaire items, is not based on a *factual* novelty as calculated by objective means.

7.5. Analysis of interaction patterns

Apart from a statistical comparison of logged interaction data, we approached our second research question (*How does the form of item space visualization influence how users interact with this visualization and the recommendations?*) also with an analysis of interaction patterns of participants, which follows the process described by Garofalakis et al. (2002) and Tsai and Brusilovsky (2019b). To achieve this, we first coded all discrete click actions according to certain categories they pertain to: E = actions in exploration area (i.e. select artist), D = actions in details area (i.e. select artist, select song, like song, unlike song), R = actions with recommendation area (i.e. select song), P = actions regarding the playlist (i.e. add to playlist, remove from playlist). In addition, we coded some auxiliary actions which only existed in our experiment: S = start task, F = finish task. We restricted the actions to those occurring during task 3 and 4 because these tasks most closely mimic a real-world situation (i.e. the creation of a playlist). Moreover, participants had access to all features of the application only during

[14] Raw log data can be found in the supplement.

Table 4
Some descriptive results of data logged during participants' interaction with the three prototypes. In case of a significant difference, the higher value is marked in bold and the condition to which the significant difference was present is indicated by a superscript letter. The number of clicks in the recommendation area differed on a significance level < .1, while the number of overall interactions and the ratio of recommendations on playlist differed on a level < .005 and < .05, respectively. Exact p-values can be found in Section 7.4.

	List		*Treemap*		*Map*	
	M	SD	M	SD	M	SD
Total completion time (minutes)	16.35	11.34	14.64	6.05	15.97	7.90
No. of clicks in exploration area	16.77	20.30	19.13	8.19	21.39	12.30
No. of clicks in details area	62.77	71.93	54.80	18.52	58.00	27.00
No. of clicks in recommendation area	12.07	8.23	8.37	4.21	**13.42**$^{\mathrm{T}}$	11.65
No. of clicks in playlist area	0.37	0.89	0.27	0.78	0.68	1.30
No. of overall interactions	163.77	208.16	**424.87**$^{\mathrm{L}}$	412.15	291.52	142.35
No. of playlist add actions	12.33	0.71	12.40	0.62	13.07	2.14
Ratio of recommendations on playlist	**0.36**$^{\mathrm{T}}$	0.29	0.18	0.16	**0.33**$^{\mathrm{T}}$	0.30
Recommendation self-information	0.70	0.10	0.73	0.15	0.71	0.13

Table 5
Top 15 interaction patterns with length ≥ 2 that we logged during task 3 and 4. The letters represent the following actions (in alphabetical order): D = interaction with details view, E = actions in exploration area, F = finish task, P = playlist add or remove, R = actions with recommendation area, S = start task.

List		*Treemap*		*Map*		Overall	
Pattern	Support	Pattern	Support	Pattern	Support	Pattern	Support
ED	90 %	DP	100 %	DP	94 %	DP	92 %
PF	87 %	PF	100 %	PF	90 %	PF	92 %
PE	87 %	ED	93 %	ED	84 %	ED	89 %
PED	87 %	EDP	93 %	EDP	84 %	PE	86 %
DP	83 %	DPF	90 %	PE	81 %	EDP	85 %
PR	83 %	PE	90 %	DPE	77 %	PED	85 %
EDP	77 %	PED	90 %	PED	77 %	DPE	78 %
PEDP	73 %	PEDP	90 %	SE	77 %	PEDP	76 %
DPE	70 %	DPE	87 %	DPF	68 %	DPF	74 %
DD	67 %	DPED	83 %	DPED	68 %	DPED	73 %
DPED	67 %	EDPE	83 %	PR	68 %	SE	70 %
RP	67 %	PD	83 %	EE	65 %	EDPE	68 %
DPF	63 %	DPEDP	77 %	PEDP	65 %	PD	67 %
PD	63 %	SE	77 %	EDPE	61 %	PR	66 %
DPEDP	60 %	DPD	73 %	EED	61 %	DPEDP	65 %
EDPE	60 %	EDPED	73 %	DD	58 %	DD	63 %

these tasks and thus were able to follow richer interaction patterns. All clicks by a single participant during each of these two tasks were interpreted as a click sequence that starts with an S and ends with an F action. For each of these sequences, we identified all subsequences with a length of at least 2 that appeared during that sequence. For each of these subsequences, we calculated its *support*. The support describes the ratio of sequences containing that subsequence proportional to the number of all sequences. Finally, we defined a support threshold of 20 %, meaning that we considered only those subsequences that we found in more than 20 % of all full sequences.

Overall, we found 151 subsequences with a support above the threshold for the *List* condition, 173 for the *Treemap* condition, and 124 for the *Map* condition. That means that in the *Treemap* condition more patterns occurred in more than 20 % of full click sequences than in the other conditions. Interactions with the other conditions thus followed a more individual style for each participant and task. A similar observation can be made for the individual support scores in Table 5: the patterns in the *Treemap* condition have fairly high support values—it is the only condition that contains patterns shared by all participants across both tasks (i.e. patterns with a support of 100 %). We note that these two patterns were an integral part of the task given to participants (*click song in details area → add song to playlist* and *add song to playlist → end task*). While these patterns also had the highest support in the *Map* condition, they were not used by all participants across both tasks (i.e. they had a support > 100 %). Thus, we had the impression that participants in the *Treemap* condition favored actions that were immediately necessary for their task.

Another aspect that becomes apparent is that in the 15 most frequent patterns in Table 5, an interaction with recommendations (i.e. patterns comprising an "R" action) occurred 2 times in the *List* condition, 1 time in the *Map* condition and 0 times in the *Treemap* condition. Obviously, recommendations were used more frequently in the *List* condition compared to the *Map* and the *Treemap* condition. Upon closer inspection, the most common pattern that included an interaction with recommendations for the *List* and *Map* conditions was "PR" (*add song to playlist → select recommended song*). In contrast, the reverse pattern ("RP") occurred less frequently. The easiest way for users to complete tasks 3 and 4 using recommendations would be a chain of six consecutive "RP". However, since "PR" occurred more frequently, this suggests that recommendations were not always added to the playlist immediately, but were examined and only used to create the playlist in some cases.

During this rather qualitative analysis of interaction pattern, we had the impression that there might be groups of participants who exhibited the same patterns regardless of the condition they were in. To explore such user groups, we applied a hierarchical clustering to the pattern data described above. Therefore, we first determined a dissimilarity score for each pair of participants based on their interaction patterns.[15] Specifically, we calculated the dissimilarity δ between a pair of participants u and v as follows:

$$\delta(u, v) = 1 - \frac{|P_u \cap P_v|}{min(|P_u|, |P_v|)},$$

where the set P_u contains all interaction patterns of participant u with a length of at least 2 from tasks 3 and 4 of our experiment.

Based on these dissimilarity scores, we performed hierarchical clustering following *Ward's agglomerative clustering method* (Ward, 1963) because it showed the best *clustering coefficient* (Kaufman and Rousseeuw, 1990) compared to other methods.[16] The result of hierarchical clustering can be seen in Fig. 11, which also shows the *cut* where we divided the cluster hierarchy. To determine this cut, we looked at the *average silhouette index* (SI) for cuts between 2 and 8 clusters. This index quantifies cohesion within clusters and separation between clusters (Rousseeuw, 1987) and ranges from −1 to 1. We found the highest SI for a cut at 2 clusters (SI = 0.042) followed by a cut at 3 clusters (SI = 0.041).[17] Consequently, we divided our dataset into two clusters: *Cluster A* (N = 34), with rather satisfied participants who were mainly in the *Map* condition, and *Cluster B* (N = 57), with "hurried" participants who were mainly in the *Treemap* condition. Below, we describe these two clusters in greater detail.

[15] For this quantitative analysis, we utilized more diverse codes than those described above, which can be found in the supplementary material.

[16] We tested four other clustering methods. All corresponding coefficients can be found in the supplementary material.

[17] Again, the results for all alternative parameters that we tested can be found in the supplement.

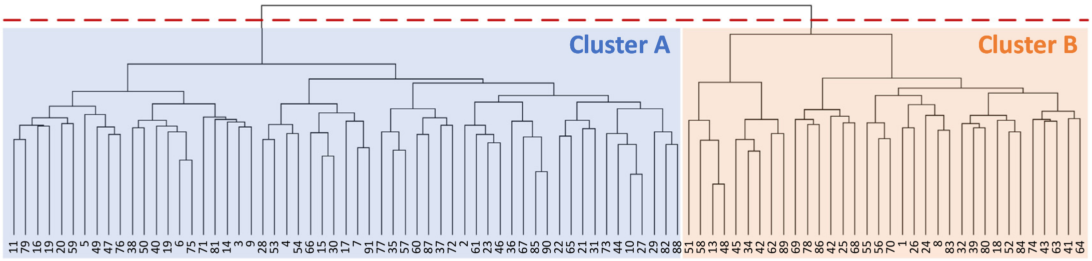


Fig. 11. A dendrogram depicting the results of clustering participants hierarchically based on the interaction patterns they used during task 3 and 4 of our experiment. The dotted line indicates the horizontal cut with the best *silhouette* value, which divides the participants into two clusters: *Cluster A* and *Cluster B*.

One of the aspects we were particularly interested in was whether the differences in participants' interaction patterns were due to a difference in the application to which they were exposed. We used a non-parametric chi-square test to compare the categorical variable *condition* between the clusters. The result showed a significant difference between the clusters ($\chi^2(2) = 5.219$, $p = .074$). While participants in the *List* condition were fairly well distributed between clusters (*Cluster A*: 32.4 %; *Cluster B*: 33.3 %), *Cluster A* had more participants of the *Map* condition (47.1 %) compared to *Cluster B* (26.3 %). Consequently, *Cluster B* contained more participants of the *Treemap* condition (40.4 %) compared to *Cluster A* (20.6 %).

To study the effects of interaction patterns on other behavioral and perceptional variables quantitatively, we conducted a MANOVA with the cluster as independent variable and the dependent variables in Tables 3, 4, and the user characteristics as introduced in Section 6.1. The result for the MANOVA was significant ($F(25, 59) = 1.684$, $p = .052$, *Wilk's* $\Lambda = .584$, $\eta_p^2 = .416$).

We analyzed between-cluster results using post hoc Bonferroni-corrected ANOVAs and found that the clusters appear to distinguish two user groups that differ mainly in their use of recommendations. For instance, we found a significant difference for the *number of clicks in the recommendation area* ($F(1, 83) = 20.388$, $p = .001$, $\eta_p^2 = .197$), with *Cluster A* grouping those participants who used the recommendation area more frequently ($M = 15.70$, $SD = 11.40$) than participants in *Cluster B* ($M = 8.11$, $SD = 3.81$). As a result, also participants' playlists in *Cluster A* consisted of a higher *ratio of recommendations* ($F(1, 83) = 15.840$, $p = .001$, $\eta_p^2 = .160$), with almost every other entry in the playlist being based on a song recommended by the system ($M = .41$, $SD = .30$). Clusters also differed in the *number of playlist add actions* ($F(1, 83) = 4.783$, $p = .032$, $\eta_p^2 = .054$), with participants in *Cluster A* adding M =13.04 (SD = 2.04) songs on average compared to participants in *Cluster B* who added M =12.35 (SD = 0.84) songs on average to their playlists. Participants also spent different times completing their tasks ($F(1, 83) = 8.296$, $p = .005$, $\eta_p^2 = .091$), with M =19.15 min (SD = 11.12) in *Cluster A* and M =13.60 min (SD = 6.67) in *Cluster B*. Apparently, this group was more satisfied with the recommendations provided by the system: during the questionnaire ($F(1, 83) = 4.865$, $p = .030$, $\eta_p^2 = .055$), participants in *Cluster A* ascribed a higher *recommendation quality* to their systems ($M = 4.26$, $SD = 0.61$) than participants in *Cluster B* ($M = 3.92$, $SD = 0.70$). Probably as a consequence, they experienced their application as of different *pragmatic* quality ($F(1, 83) = 3.864$, $p = .053$, $\eta_p^2 = .044$), with higher scores in *Cluster A* ($M = 5.30$, $SD = 1.18$) compared to *Cluster B* ($M = 4.67$, $SD = 1.53$).

7.6. Influence of user characteristics

To tackle our third research question (*How do user characteristics influence perceptions of item space and the recommendations and how users interact with them?*), we calculated two-tailed Pearson correlation coefficients to assess the linear relationship between the dependent variables listed in Tables 3 and 4, and the user characteristics we elicited as part of the initial questionnaire (Section 6.1).

Overall, we only found few significant correlations. Participants who scored high on the dimension of *neuroticism* of the Big Five perceived a lower *recommendation variety* ($r(91) = -.182$, $p = .078$). Those scoring high on *openness*, however, perceived a higher *recommendation transparency* ($r(91) = .222$, $p = .035$). This dimension surprisingly also correlated with the recommendation *self-information* ($r(91) = .245$, $p = .019$), which we calculated as objective measurement of recommendation novelty (Eq. (9)). It seems that open-minded participants received more novel recommendations (but did not perceive them as such). We also found a positive correlation with this recommendation *self-information* and a *rational decision style* ($r(91) = .187$, $p = .076$). Participants following this style of decision making also perceived their recommendations as more transparent ($r(91) = .204$, $p = .053$). In contrast, a *rational decision style* correlated negatively with the *ratio of recommendation on playlist* ($r(91) = -.207$, $p = .049$). Participants that base their decisions on rational reasons, used less recommendations for their playlist. Apart from that, we did not find any significant correlations of the user characteristics we elicited and our dependent variables.

With respect to differences between clusters as described in the section above, we found that clusters differed significantly in terms of *agreeableness* ($F(1, 83) = 5.328$, $p = .023$, $\eta_p^2 = .060$) with a higher score on that dimensions in *Cluster A* ($M = 4.01$, $SD = 0.55$) compared to *Cluster B* ($M = 3.70$, $SD = 0.61$). Apart from that, participants in *Cluster A* had with M =6.17 (1.23) a higher capacity of their *visual memory* ($F(1, 83) = 3.543$, $p = .063$, $\eta_p^2 = .041$) compared to participants in *Cluster B* ($M = 5.58$, $SD = 1.44$).

8. Discussion

Our overarching goal with this research was to study effects of the visualization of the item space on perceptional aspects of recommendations (i.e. perceived transparency, quality, variety, and novelty of recommendations) and on the general user experience. In this chapter, we discuss our research questions in light of the results of our study and evaluate them in term of achievements and limitations.

8.1. RQ1: How does the type of item space visualization influence users' perception and understanding of the item domain and the recommendations?

Previous work has pointed to several advantages of using InfoVis methods in RS. It was, for example, observed that InfoVis may increase perceived recommendation transparency, diversity, and control. However, in this experiment, we were not able to replicate these findings. We see two general ways to interpret this: (1) Our two InfoVis interfaces (*Treemap* and *Map*) might not have been designed properly to improve such perceptual variables over the baseline *List* interface. (2) The previously reported effects are generally overestimated. While we acknowledge that we cannot provide a definitive answer to this question, we argue for the second interpretation. The main argument for this perspective is that baseline systems used in prior studies did not include appropriate means for browsing the item space. We thus argue that even simple browsing functions like those provided in our *List* prototype together with a depiction of preferences and recommendations close to each other might be sufficient to provide an adequate level of perceived transparency, overview, and control in RS. Especially in terms of *recommendation transparency*, our results suggest that participants in all three conditions experienced their recommendations as highly transparent (i.e. the transparency was rated as equal to or above 4 on a 5-point Likert scale, see Table 3). We attribute these high transparency values of our results to the visual proximity of preferences and recommendations as visible in the headers (*List* condition), in the top rows of each genre (*Treemap* condition), and in the distance between artist locations (*Map* condition). Preferences and recommendations might have created something like an "explanatory proximity", which is probably as intuitively comprehensible as methods strictly based on the *first law of geography* discussed in Section 2.2.

This is particularly noteworthy because we elicited the *perceived* transparency, and thus the transparency of recommendations experienced by users, not actual transparency, for which we would have needed to provide real insights into the recommendation algorithm, which we did not have access to in this experiment. At this point, we suggest to further investigate the differences between perceived and actual transparency in future research. Previous studies indicate that when the perception of how an intelligent system works is too different from how it actually works, it can lead to different problems in human–computer interaction (Ananny and Crawford, 2018; Kunkel et al., 2021).

In contrast to transparency, user experience is a dimension which differed significantly. More specifically, we found that the *pragmatic user experience* was significantly higher in the *List* condition than in the *Treemap* condition, and that the *hedonic user experience* was significantly higher in the *Map* condition in comparison with both, the *List* and the *Treemap* condition. Apparently, participants in the *Map* condition were able to navigate a space with a large number of simultaneously presented artists. In other words, the user experience was not linked to the number of information objects displayed. The scores for visual clutter as measured by SE (see Section 5.4) followed this trend as well and thus are not reliable predictors of user experience either. However, we also measured FC, which indicates more clutter in the *Treemap* condition. We attribute this to the use of many similar visual objects (artist labels) that are relatively evenly distributed vertically and horizontally throughout the visualization. Therefore, it would not be so easy to "add a new item that would reliably draw attention" (Rosenholtz et al., 2007) in the *Treemap* interface. As such, FC takes into account the orientation and organization of objects, e.g. whether similar objects are grouped together, which was apparently more the case in the *Map* interface—probably due to visualizing artists as clusters. So, regardless of the reasons, FC seems to be the more appropriate measurement to indicate cluttered displays and, as a consequence, to predict user experience. We conclude, that distributing the artist in our *Map* condition based on their similarity seems to be an effective way to use the available space of a two-dimensional screen. In fact, the *Map* was able to increase the number of simultaneously displayed information objects by a factor of over seven with respect to the *List* condition and a factor of about three with respect to the *Treemap* condition, while at the same time increasing the hedonic user experience. With this application, we hence demonstrate how it is possible to make the exploration of large music datasets more efficient and entertaining as, for instance, been called for by Schedl (2017).

Apart from the *hedonic user experience*, we also noted that the perceived *recommendation novelty* was higher in the *Map* condition than in the other two conditions. Backed up by a mediation analysis, we argue that this may be due to a *halo effect*. The halo effect describes a cognitive error in which characteristics are attributed to a person based on their attractiveness rather than on a factual perception of those attributed characteristics (Thorndike, 1920; Nisbett and Wilson, 1977). Hassenzahl and Monk (2010) argue for a similar effect in web design, showing that the perceived beauty of a website affects its perceived usability. We transfer this to the hedonic user experience of our web application and the perceived novelty of recommendations. To our knowledge, this paper is the first to suggest such a relationship. Yet, the effect itself may have been observed earlier (e.g. in the experiment of Millecamp et al. (2018), in which a higher perceived novelty of recommendations was measured in the experimental condition as well). However, we acknowledge that this relation is rather speculative and demands further studies. It is likely that a halo effect of perceived innovativeness of an application may have an affect on how recommendations are perceived, but not necessarily. In reality there likely is a mix of different effects. Either way, we suggest looking at the evaluation of RS from a more holistic perspective, including thus far underexplored side-effects of user experience on other user-centered criteria.

When analyzing the qualitative comments, we found that many participants rated the quality of the recommendations during the experiment higher compared to their usual Spotify experience. This is especially noteworthy given that we relied on the recommendations provided by the Spotify API. We thus expected either no difference or one in favor of the recommendations at Spotify, as these are likely to be based on a substantially larger dataset of preferences. One factor that could account for this surprising difference in perceived recommendation quality is the observation that the diversity of recommendations decreases over time (Nguyen et al., 2014). This, in turn, may have led participants to perceive the music recommendations in our experiment as surprisingly different from those they normally receive. As a consequence, we suggest to provide an option to users to temporarily receive recommendations that are not as personalized as they typically are—or in other words, to temporarily leave their filter bubble. Another factor that may have influenced perceived recommendation quality is that participants rated the recommendations as highly transparent. A relationship between transparency and recommendation quality has been reported several times in the past (Bilgic and Mooney, 2005; Donkers et al., 2016; Kunkel et al., 2019). While we have no way of knowing whether the Spotify's transparency would be judged to be lower, the lack of the "explanatory proximity" we mentioned above at Spotify,[18] suggests that it would be.

8.2. *RQ2: How does the type of item space visualization influence how users interact with this visualization and the recommendations?*

The comparison of different log data variables reveal that participants in the *Treemap* condition made considerable less use of recommendations as observed in a lower *number of clicks in the recommendation area* and a lower *ratio of recommendations* on the participants' final playlist during task 3 and 4. The difference in the *number of overall interactions* but not in the number of clicks, indicates, that participants in the *Treemap* condition used more scroll actions than participants in the *List* condition. This is especially noteworthy since in both conditions, scrolling was used to browse the list of artists in a subgenre. One reason for this difference may be that, due to space constraint, sample artists, currently selected artists, and artists with preferred or recommended songs were depicted on a vertical axis and therefore required more scroll actions to browse the space of artists. The visual separation of this special list of artists and the rest of the artists may have been perceived as cognitively easier to comprehend in the *List* condition—probably due to the formation of visual and conceptual unity supported by *Gestalt laws* such as *proximity* and *common region*.

The observation of a difference in the use of recommendations between conditions can also be found when clustering participants regarding their interaction patterns. In doing so, we found that the way users perceive and interact with an application is only partially

[18] The only parts where we know of a link between preferences and recommendations in Spotify are the "play song radio" and "play artist radio" options.

dependent on the application presented to them. Instead, our clustering suggests that one group of participants used the recommendations significantly more often and perceived them to be of higher quality than the rest of the sample. This group (*Cluster A*) consisted of participants in the *Map* or *List* condition, less so of participants in the *Treemap* condition, which may also explain the difference between conditions regarding the use of recommendations. Apart from the usage and the quality perception of recommendations, participants in *Cluster A* experienced several other aspects of the applications as high, too. This can probably partly be ascribed to the fact that participants in *Cluster A* scored higher on the Big Five dimension of *agreeableness*, which may indicate that they are less critical when assessing the interface and recommendations. On the other hand, participants in this group also had a higher *visual memory* capacity, which probably helped them in cognitively processing the information objects presented to them. However, since there were few direct correlations between user characteristics and the dependent perceptional and behavioral variables, the differences measured between clusters appear to be due to another factor. One possible candidate for this is that the clustering separated participants who held a different *mental model*. In an earlier experiment, we found that such mental models can indeed be responsible for perceptions of RS (Kunkel et al., 2021). We therefore suggest that future work should focus on such complex interactions between user characteristics, mental models, and perceptual variables. In any case, we consider this as another argument for taking a more holistic perspective when evaluating RS and InfoVis applications rather than treating the individual aspects separately.

8.3. RQ3: How do users' characteristics influence how they perceive, understand, and interact with the item space and the recommendations?

In regards of such a *holistic* perspective, an ongoing line of research investigates the influence of personality traits and user characteristics in RS (e.g. Tkalcic and Chen, 2015; Millecamp et al., 2018; Jin et al., 2019) and InfoVis (e.g. Ziemkiewicz et al., 2011; Conati et al., 2015; Lallé and Conati, 2019). It was found that the perception of RS and InfoVis applications may well be influenced by certain user characteristics. We, however, did not find many of such relationships. We were particularly surprised that the capacity of participants' *visual memory* did not influence perceptions of the item space, for instance, the perceived *degree of overview*. We conclude that further work needs to be conducted to investigate in greater detail when such effects arise and when they do not. We also acknowledge that there may be other factors, which may have influenced the results but which we did not measure. Examples include *musical sophistication* (Müllensiefen et al., 2014) and *locus of control* (Green and Fisher, 2010), which have been found to influence how interactive controls of RS are used (Millecamp et al., 2018) and the performance of interacting with InfoVis applications (Ziemkiewicz et al., 2011), respectively. Both user characteristics are candidates for investigating their influence on how item space visualizations are perceived by users in future work.

However, the fact that we only found few correlations between user characteristics and dependent variables of our experiment suggests that the role of such characteristics demands further discussion. We hence conclude that future research needs to investigate *why* and *when* user characteristics influence perceptions of RS and InfoVis applications. One tool that is particularly suited to answer such *why* questions is to conduct exploratory qualitative studies. By following the method of *grounded theory* (Strauss and Corbin, 1994; Corbin and Strauss, 2008), for instance, researchers could leverage *theoretical sampling* and specifically sample participants with a particular user characteristic to develop a theory about why and under which circumstances this characteristic influences the perception of recommendations or of the visualization used. This could possibly also reveal the mental models we mentioned above. In fact, we used this method in an exploratory study before (Ngo et al., 2020).

8.4. Limitations

We decided to conduct our study as an online study, due to local Covid-19 restrictions. For our use case an online study was not ideal, though. First, because we were not able to control confounding variables as well as we could in a lab setting. One of these confounding factors is the device that participants used to access the web application. For example, we know that the *Treemap* and *Map* interfaces require at least 8 GB RAM to function smoothly. As a result, we received comments mentioning that the application was "laggy" or "stuttery". During the experiment, we also received some direct messages from participants pointing out technical issues. In most cases, such participants were using exotic browser versions or content-blocking addons. A second limitation due to conducting the study online was that some participants finished the experiment relatively quickly (the fastest completed the study in about 12 min, while test candidates in a brief prestudy took about 45 min on average, which was also the value we used to calculate monetary compensation for crowdworkers). In lab studies, on the other hand, we found that participants put more effort into completing their tasks, probably because the situation seems more serious or official to them. Although these circumstances may have affected the quality of the user experience, participants were randomly assigned to a condition, and we are thus confident that potential negative effects did not affect our experiment too much.

Another limitation concerns the sample size, which is rather small—especially for an online experiment. This was due to resource limitations and the requirements we set up for accepting participants. Still, the sample allowed us to obtain significant results, many at the lower .05 alpha level. However, we conclude that further experiments should be conducted to corroborate our findings. To this end, we added a supplementary file archive to this article that provides additional details on our study that, among other things, could help researchers build upon our results and further advance the field of user-centered RS and InfoVis research. In this context, we would also like to add that we deliberately decided for the music domain, but suggest that future research should also take other domains into account. It would be particularly interesting to see whether the halo effect, which we suspect to have influenced the perception of recommendation novelty, also occurs for *search products* or in domains with higher risk involved.

9. Conclusion and future work

In this paper, we introduce three applications that each take their own perspective on the same dataset: a hierarchically structured set of musical artists. Among typical interactions to explore the dataset such as searching, browsing, and examining single songs, users can also rate items and receive matching recommendations. A user study revealed that there are no differences between conditions in the main perceptive dimensions of Recommender Systems, such as transparency, recommendation quality, and degree of control. However, we found that the three applications differed significantly in the user experience, especially in the hedonic dimension. Also the perceived novelty of recommendations differed significantly between the three applications, even though an objective difference was not found. We thus conclude that a halo effect might be at work here: the *Map* condition was perceived as innovative and leading edge, which in turn let participants experience its recommendations as more novel. A conducted mediation analysis backs up this hypothesis and shows that the effect of condition on perceived novelty is completely mediated by hedonic user experience.

We acknowledge that these findings are still speculative to some degree and conclude that such halo effects in RS and InfoVis should be studied in greater depth in the future especially by experiments isolating the effect. With regards to advancing the applications, we consider to explore more aspects of the *Map* visualization. One candidate aspect for improvements is to conduct studies on how the item space abstraction and sample selection should be performed. In our setting, there were two representative sample items for each genre and subgenre. It remains an open question, though, whether a setting with more samples per area would allow users to more easily grasp the underlying space. On the other hand, this might also lead to cognitive overload, so in reality a compromise has to be found.

Another aspect which we deem interesting to further investigate is to use maps for dynamically visualizing trends (e.g. in terms of changing user preferences). This is especially relevant in the domain of music, where the preferences can be rather ephemeral and change due to mood, activities, or other contextual conditions. For this, further geographic metaphors could be exploited. It would, for instance, be imaginable to first record how preferences change over the course of a day (e.g. by logging listening behavior without the use of a map). In a daily summary, one could project these changing preference by animating a layer of the map resembling a weather report. This could also let users interactively set their music recommendations to a certain *weather condition* to easily tell the system which music the want to listen to at the moment. Also in general, the layout of the *Map* application could follow a more realistic layout, for example by including element of geographic maps such as rivers, cities, or roads. This was suggested by our participants in their qualitative answers and is in line with existing evidence (Montello et al., 2003; Pang et al., 2016, 2017).

CRediT authorship contribution statement

Johannes Kunkel: Conceptualization, Methodology, Implementation, Execution of online study, Manuscript draft. **Jürgen Ziegler:** Supervision, Finalization of manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Supplementary material is available at https://data.mendeley.com/datasets/hw39d5kxs6/3.

References

Ananny, M., Crawford, K., 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. 20 (3), 973–989.

Anantharam, V., Varaiya, P., Walrand, J., 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards. IEEE Trans. Automat. Control 32 (11), 968–976. http://dx.doi.org/10.1109/TAC.1987.1104491.

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., Lalmas, M., 2020. Algorithmic effects on the diversity of consumption on spotify. In: Proceedings of the Web Conference 2020. WWW '20, Association for Computing Machinery, pp. 2155–2165. http://dx.doi.org/10.1145/3366423.3380281.

Andjelkovic, I., Parra, D., O'Donovan, J., 2019. Moodplay: Interactive music recommendation based on Artists' mood similarity. Int. J. Hum.-Comput. Stud. 121, 142–159. http://dx.doi.org/10.1016/j.ijhcs.2018.04.004.

Bakshy, E., Messing, S., Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348 (6239), 1130–1132. http://dx.doi.org/10.1126/science.aaa1160.

Bendada, W., Salha, G., Bontempelli, T., 2020. Carousel personalization in music streaming apps with contextual bandits. In: Fourteenth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, pp. 420–425. http://dx.doi.org/10.1145/3383313.3412217.

Bilgic, M., Mooney, R.J., 2005. Explaining recommendations: satisfaction vs. promotion. In: Proceedings of beyond Personalization Workshop, IUI.

Biuk-Aghai, R.P., Pang, P.C.-I., Pang, B., 2017. Map-like Visualisations vs. Treemaps: An experimental comparison. In: Proceedings of the 10th International Symposium on Visual Information Communication and Interaction. VINCI '17, ACM, pp. 113–120. http://dx.doi.org/10.1145/3105971.3105976.

Biuk-Aghai, R.P., Pang, C.-I., Si, Y.-W., 2014. Visualizing large-scale human collaboration in Wikipedia. Future Gener. Comput. Syst. 31, 120–133. http://dx.doi.org/10.1016/j.future.2013.04.001.

Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling: Theory and Applications, second ed. Springer, New York.

Boy, J., Rensink, R.A., Bertini, E., Fekete, J.-D., 2014. A principled way of assessing visualization literacy. IEEE Trans. Vis. Comput. Graphics 20 (12), 1963–1972. http://dx.doi.org/10.1109/TVCG.2014.2346984.

Bruls, M., Huizing, K., van Wijk, J.J., 2000. Squarified treemaps. In: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Amsterdam. In: Data Visualizaion 2000, Springer Vienna, pp. 33–42. http://dx.doi.org/10.1007/978-3-7091-6783-0_4.

Burke, R., Sonboli, N., Ordonez-Gauger, A., 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In: Friedler, S.A., Wilson, C. (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency. In: Proceedings of Machine Learning Research, vol. 81, PMLR, pp. 202–214, URL: https://proceedings.mlr.press/v81/burke18a.html.

Card, S.K., Mackinlay, J.D., Shneiderman, B., 1999. Readings in Information Visualization: Using Vision to Think. In: The Morgan Kaufmann Series in Interactive Technologies, Morgan Kaufmann Publishers, San Francisco, Calif.

Cardoso, B., Sedrakyan, G., Gutiérrez, F., Parra, D., Brusilovsky, P., Verbert, K., 2019. IntersectionExplorer, a multi-perspective approach for exploring recommendations. Int. J. Hum.-Comput. Stud. 121, 73–92, URL: https://www.sciencedirect.com/science/article/pii/S1071581918301903.

Chang, J.C., Hahn, N., Perer, A., Kittur, A., 2019. SearchLens: Composing and capturing complex user interests for exploratory search. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19, ACM, pp. 498–509, URL: https://doi.org/10.1145/3301275.3302321.

Chen, L., Tsoi, H.K., 2011. Users' decision behavior in recommender interfaces: Impact of layout design. In: RecSys' 11 Workshop on Human Decision Making in Recommender Systems.

Conati, C., Carenini, G., Hoque, E., Steichen, B., Toker, D., 2014. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. Comput. Graph. Forum 33 (3), 371–380. http://dx.doi.org/10.1111/cgf.12393.

Conati, C., Carenini, G., Toker, D., Lallé, S., 2015. Towards user-adaptive information visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI '15, AAAI Press, pp. 4100–4106.

Corbin, J., Strauss, A., 2008. Basics of Qualitative Research (3rd Ed.): Techniques and Procedures for Developing Grounded Theory. SAGE Publications, Thousand Oaks, California.

Corsi, P.M., 1972. Human Memory and the Medial Temporal Region of the Brain (Ph.D. thesis). McGill University, Montreal, Canada.

Cunningham, S.J., Reeves, N., Britland, M., 2003. An ethnographic study of music information seeking: Implications for the design of a music digital library. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '03, IEEE Computer Society, pp. 5–16, URL: http://dl.acm.org/citation.cfm?id=827140.827142.

Dandekar, P., Goel, A., Lee, D.T., 2013. Biased assimilation, homophily, and the dynamics of polarization. Proc. Natl. Acad. Sci. 110 (15), 5791–5796. http://dx.doi.org/10.1073/pnas.1217220110.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W., 2016. The spreading of misinformation online. Proc. Natl. Acad. Sci. USA 113 (3), 554–559. http://dx.doi.org/10.1073/pnas.1517441113.

DiFranzo, D., Gloria-Garcia, K., 2017. Filter bubbles and fake news. XRDS 23 (3), 32–35. http://dx.doi.org/10.1145/3055153.

Donkers, T., Loepp, B., Ziegler, J., 2016. Tag-enhanced collaborative filtering for increasing transparency and interactive control. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. UMAP '16, ACM, pp. 169–173. http://dx.doi.org/10.1145/2930238.2930287.

Du, F., Malik, S., Theocharous, G., Koh, E., 2018. Personalizable and interactive sequence recommender system. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. In: CHI EA '18, Association for Computing Machinery, pp. 1–6. http://dx.doi.org/10.1145/3170427.3188506.

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., Vonderau, P., 2019. Spotify Teardown : Inside the Black Box of Streaming Music.

Fabrikant, S.I., Montello, D.R., Mark, D.M., 2006. The distance-similarity metaphor in region-display spatializations. IEEE Comput. Graph. Appl. 26 (4), 34–44. http://dx.doi.org/10.1109/MCG.2006.90.

Faridani, S., Bitton, E., Ryokai, K., Goldberg, K., 2010. Opinion space: A scalable tool for browsing online comments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10, ACM, pp. 1175–1184.

Ferwerda, B., Yang, E., Schedl, M., Tkalcic, M., 2015. Personality traits predict music taxonomy preferences. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '15, Association for Computing Machinery, pp. 2241–2246. http://dx.doi.org/10.1145/2702613.2732754.

Flaxman, S., Goel, S., Rao, J.M., 2016. Filter bubbles, echo chambers, and online news consumption. Public Opin. Q. 80 (S1), 298–320. http://dx.doi.org/10.1093/poq/nfw006.

Gansner, E., Hu, Y., Kobourov, S., 2010. GMap: Visualizing graphs and clusters as maps. In: Visualization Symposium, Pacific Asia-Pacific. pp. 201–208.

Gärdenfors, P., 2004. Conceptual Spaces: The Geometry of Thought. MIT Press.

Garofalakis, M., Rastogi, R., Shim, K., 2002. Mining sequential patterns with regular expression constraints. IEEE Trans. Knowl. Data Eng. 14 (3), 530–552. http://dx.doi.org/10.1109/TKDE.2002.1000341.

Giboney, J.S., Brown, S.A., Lowry, P.B., Nunamaker, J.F., 2015. User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. Decis. Support Syst. 72, 1–10. http://dx.doi.org/10.1016/j.dss.2015.02.005.

Green, T.M., Fisher, B., 2010. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In: 2010 IEEE Symposium on Visual Analytics Science and Technology. pp. 203–210. http://dx.doi.org/10.1109/VAST.2010.5653587.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Hall, C., Höllerer, T., 2010. SmallWorlds: Visualizing social recommendations. Comput. Graph. Forum 29 (3), 833–842. http://dx.doi.org/10.1111/j.1467-8659.2009.01679.x.

Gunawardana, A., Shani, G., 2015. Evaluating recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. Springer US, Boston, MA, pp. 265–308.

Haim, M., Graefe, A., Brosius, H.-B., 2018. Burst of the filter bubble? Effects of personalization on the diversity of google news. Digit. Journal. 6 (3), 330–343. http://dx.doi.org/10.1080/21670811.2017.1338145.

Hamilton, K., Shih, S.-I., Mohammed, S., 2016. The development and validation of the rational and intuitive decision styles scale. J. Personal. Assess. 98 (5), 523–535. http://dx.doi.org/10.1080/00223891.2015.1132426.

Hassenzahl, M., Monk, A., 2010. The inference of perceived usability from beauty. Hum.–Comput. Interact. 25 (3), 235–260. http://dx.doi.org/10.1080/07370024.2010.500139.

Hayes, A.F., 2022. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, third ed. In: Methodology in the Social Sciences, Guilford Press, New York, NY, US.

He, C., Parra, D., Verbert, K., 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. Expert Syst. Appl. 56, 9–27. http://dx.doi.org/10.1016/j.eswa.2016.02.013.

Hellman, M., Hernandez-Bocanegra, D.C., Ziegler, J., 2022. Development of an instrument for measuring users' perception of transparency in recommender systems 156-165. In: Smith-Renner, A., Amir, O. (Eds.), Joint Proceedings of the IUI 2022 Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE Co-Located with the ACM International Conference on Intelligent User Interfaces (IUI 2022), Virtual Event, Helsinki, Finland, March 21-22, 2022. In: CEUR Workshop Proceedings, vol. 3124, CEUR-WS.org, pp. 156–165, URL: http://ceur-ws.org/Vol-3124/paper17.pdf.

Herlocker, J.L., Konstan, J.A., Riedl, J., 2000. Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. CSCW '00, ACM, pp. 241–250. http://dx.doi.org/10.1145/358916.358995.

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J., 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22 (1), 5–53.

Hurley, N., Zhang, M., 2011. Novelty and diversity in top-n recommendation – analysis and evaluation. ACM Trans. Internet Technol. 10 (4), http://dx.doi.org/10.1145/1944339.1944341.

Iyengar, S.S., Lepper, M.R., 2000. When choice is demotivating: Can one desire too much of a good thing? J. Personal. Soc. Psychol. 76 (6), 995–1006. http://dx.doi.org/10.1037/0022-3514.79.6.995.

Jannach, D., Adomavicius, G., 2016. Recommendations with a purpose. In: Proceedings of the 10th ACM Conference on Recommender Systems. RecSys '16, ACM, pp. 7–10. http://dx.doi.org/10.1145/2959100.2959186.

Jin, Y., Tintarev, N., Htun, N.N., Verbert, K., 2019. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. User Model. User-Adapt. Interact. (30), 199–249. http://dx.doi.org/10.1007/s11257-019-09247-2.

Kardes, F.R., Posavac, S.S., Cronley, M.L., 2004. Consumer inference: A review of processes, bases, and judgment contexts. J. Consum. Psychol. 14 (3), 230–256. http://dx.doi.org/10.1207/s15327663jcp1403_6.

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., Castrén, E., 2003. Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics 4 (1), 48. http://dx.doi.org/10.1186/1471-2105-4-48.

Katarya, R., Jain, I., Hasija, H., 2014. An interactive interface for instilling trust and providing diverse recommendations. In: 2014 International Conference on Computer and Communication Technology. ICCCT, pp. 17–22. http://dx.doi.org/10.1109/ICCCT.2014.7001463.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. In: Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ.

Kayo, O., 2006. Locally Linear Embedding Algorithm: Extensions and Applications (Ph.D. thesis). University of Oulu, URL: http://herkules.oulu.fi/isbn9514280415/.

Knees, P., Schedl, M., Masataka, G., 2019. Intelligent user interfaces for music discovery: The past 20 years and what's to come. In: 20th International Society for Music Information Retrieval Conference. In: ISMIR 2019.

Knees, P., Schedl, M., Pohle, T., Widmer, G., 2006. An innovative three-dimensional user interface for exploring music collections enriched. In: Proceedings of the 14th ACM International Conference on Multimedia. MM '06, ACM, pp. 17–24. http://dx.doi.org/10.1145/1180639.1180652.

Knijnenburg, B.P., Sivakumar, S., Wilkinson, D., 2016. Recommender systems for self-actualization. In: Proceedings of the 10th ACM Conference on Recommender Systems. RecSys '16, ACM, pp. 11–14. http://dx.doi.org/10.1145/2959100.2959189.

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C., 2012. Explaining the user experience of recommender systems. User Model. User-Adapt. Interact. 22 (4), 441–504. http://dx.doi.org/10.1007/s11257-011-9118-4.

Konstan, J.A., Riedl, J., 2012. Recommender systems: from algorithms to user experience. User Model. User-Adapt. Interact. 22 (1–2), 101–123. http://dx.doi.org/10.1007/s11257-011-9112-x.

Kuhn, W., Blumenthal, B., 1996. Spatialization: Spatial metaphors for user interfaces. In: Conference Companion on Human Factors in Computing Systems. CHI '96, ACM, pp. 346–347. http://dx.doi.org/10.1145/257089.257361.

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., Ziegler, J., 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, ACM, pp. 1–12. http://dx.doi.org/10.1145/3290605.3300717.

Kunkel, J., Loepp, B., Ziegler, J., 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, pp. 3–15. http://dx.doi.org/10.1145/3025171.3025189.

Kunkel, J., Ngo, T., Ziegler, J., Krämer, N., 2021. Identifying group-specific mental models of recommender systems: A novel quantitative approach. In: Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., Inkpen, K. (Eds.), Human-Computer Interaction – INTERACT 2021. Springer International Publishing, pp. 383–404. http://dx.doi.org/10.1007/978-3-030-85610-6_23.

Kunkel, J., Schwenger, C., Ziegler, J., 2020. NewsViz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20, Association for Computing Machinery, pp. 126–135. http://dx.doi.org/10.1145/3340631.3394869.

Lallé, S., Conati, C., 2019. The role of user differences in customization: A case study in personalization for infovis-based content. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19, ACM, pp. 329–339. http://dx.doi.org/10.1145/3301275.3302283.

Lee, J.H., Downie, J.S., 2004. Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In: Proceedings of the 5th International Conference on Music Information Retrieval. ISMIR '04.

Liang, Y., Willemsen, M.C., 2021. Interactive music genre exploration with visualization and mood control. In: 26th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, pp. 175–185.

Lindgaard, G., Fernandes, G., Dudek, C., Brown, J., 2006. Attention web designers: You have 50 milliseconds to make a good first impression!. Behav. Inf. Technol. 25 (2), 115–126. http://dx.doi.org/10.1080/01449290500330448.

Loepp, B., Donkers, T., Kleemann, T., Ziegler, J., 2018. Impact of item consumption on assessment of recommendations in user studies. In: Proceedings of the 12th ACM Conference on Recommender Systems. ACM, pp. 49–53. http://dx.doi.org/10.1145/3240323.3240375.

Ma, B., Lu, M., Taniguchi, Y., Konomi, S., 2021. CourseQ: the impact of visual and interactive course recommendation in university environments. Res. Pract. Technol. Enhanc. Learn. 16 (1), 18. http://dx.doi.org/10.1186/s41039-021-00167-7.

McCrae, R.R., John, O.P., 1992. An introduction to the five-factor model and its applications. J. Personal. 60 (2), 175–215. http://dx.doi.org/10.1111/j.1467-6494.1992.tb00970.x.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., Mehrotra, R., 2018. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In: Proceedings of the 12th ACM Conference on Recommender Systems. ACM, pp. 31–39. http://dx.doi.org/10.1145/3240323.3240354.

McNee, S.M., Riedl, J., Konstan, J.A., 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: CHI '06 Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '06, ACM, pp. 1097–1101. http://dx.doi.org/10.1145/1125451.1125659.

Meinecke, C., Hakimi, A.D., Jänicke, S., 2022. Explorative visual analysis of rap music. Information 13 (1), http://dx.doi.org/10.3390/info13010010.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space.

Millecamp, M., Htun, N.N., Conati, C., Verbert, K., 2020. What's in a user? Towards personalising transparency for music recommender interfaces. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20, Association for Computing Machinery, pp. 173–182.

Millecamp, M., Htun, N.N., Jin, Y., Verbert, K., 2018. Controlling spotify recommendations: Effects of personal characteristics on music recommender user interfaces. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. UMAP '18, ACM, pp. 101–109. http://dx.doi.org/10.1145/3209219.3209223.

Montello, D.R., Fabrikant, S.I., Ruocco, M., Middleton, R.S., 2003. Testing the first law of cognitive geography on point-display spatializations. In: Kuhn, W., Worboys, M.F., Timpf, S. (Eds.), Spatial Information Theory. Foundations of Geographic Information Science. Springer Berlin Heidelberg, pp. 316–331. http://dx.doi.org/10.1007/978-3-540-39923-0_21.

Müllensiefen, D., Gingras, B., Musil, J., Stewart, L., 2014. The musicality of non-musicians: An index for assessing musical sophistication in the general population. PLoS ONE 9 (2), http://dx.doi.org/10.1371/journal.pone.0089642.

Nagulendra, S., Vassileva, J., 2014. Understanding and controlling the filter bubble through interactive visualization: A user study. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media. HT '14, ACM, pp. 107–115. http://dx.doi.org/10.1145/2631775.2631811.

Ngo, T., Kunkel, J., Ziegler, J., 2020. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, pp. 183–191. http://dx.doi.org/10.1145/3340631.3394841.

Nguyen, T.T., Hui, P.-M., Harper, F.M., Terveen, L., Konstan, J.A., 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14, ACM, pp. 677–686. http://dx.doi.org/10.1145/2566486.2568012.

Nisbett, R.E., Wilson, T.D., 1977. The halo effect: Evidence for unconscious alteration of judgments. J. Personal. Soc. Psychol. 35 (4), 250–256. http://dx.doi.org/10.1037/0022-3514.35.4.250.

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., Höllerer, T., 2008. PeerChooser: Visual interactive recommendation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08, ACM, pp. 1085–1088. http://dx.doi.org/10.1145/1357054.1357222.

Pachet, F., Cazaly, D., 2000. A taxonomy of musical genres. In: Content-Based Multimedia Information Access - Volume 2. RIAO '00, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 1238–1245.

Pampalk, E., Rauber, A., Merkl, D., 2002. Content-based organization and visualization of music archives. In: Proceedings of the Tenth ACM International Conference on Multimedia. MULTIMEDIA '02, Association for Computing Machinery, pp. 570–579. http://dx.doi.org/10.1145/641007.641121.

Pang, P.C.-I., Biuk-Aghai, R.P., Yang, M., 2016. What makes you think this is a map? Suggestions for creating map-like visualisations. In: Proceedings of the 9th International Symposium on Visual Information Communication and Interaction. VINCI '16, ACM, pp. 75–82. http://dx.doi.org/10.1145/2968220.2968239.

Pang, P.C.-I., Biuk-Aghai, R.P., Yang, M., Pang, B., 2017. Creating realistic map-like visualisations: Results from user studies. J. Vis. Lang. Comput. 43, 60–70. http://dx.doi.org/10.1016/j.jvlc.2017.09.002.

Pariser, E., 2011. The Filter Bubble: What the Internet is Hiding from You. The Penguin Press, New York, NY, USA.

Parra, D., Brusilovsky, P., Trattner, C., 2014. See what you want to see: visual user-driven approach for hybrid recommendation. In: Proceedings of the 19th International Conference on Intelligent User Interfaces. IUI '14, pp. 235–240. http://dx.doi.org/10.1145/2557500.2557542.

Petridis, S., Daskalova, N., Mennicken, S., Way, S.F., Lamere, P., Thom, J., 2022. TastePaths: Enabling deeper exploration and understanding of personal preferences in recommender systems. In: 27th International Conference on Intelligent User Interfaces. IUI '22, Association for Computing Machinery, pp. 120–133. http://dx.doi.org/10.1145/3490099.3511156.

Pu, P., Chen, L., Hu, R., 2011. A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11, ACM, pp. 157–164. http://dx.doi.org/10.1145/2043932.2043962.

Rammstedt, B., John, O.P., 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. J. Res. Personal. 41 (1), 203–212. http://dx.doi.org/10.1016/j.jrp.2006.02.001.

Rawlings, D., Ciancarelli, V., 1997. Music preference and the five-factor model of the NEO personality inventory. Psychol. Music 25 (2), 120–132. http://dx.doi.org/10.1177/0305735697252003.

Ricci, F., Rokach, L., Shapira, B. (Eds.), 2022. Recommender Systems Handbook. Springer US, New York, NY.

Richthammer, C., Pernul, G., 2017. Explorative analysis of recommendations through interactive visualization. In: E-Commerce and Web Technologies. In: EC-Web 2017, Springer International Publishing, pp. 46–57. http://dx.doi.org/10.1007/978-3-319-53676-7_4.

Rosenholtz, R., Li, Y., Nakano, L., 2007. Measuring visual clutter. J. Vis. 7 (2), 17. http://dx.doi.org/10.1167/7.2.17.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Roy, Q., Zhang, F., Vogel, D., 2019. Automation accuracy is good, but high controllability may be better. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, ACM, pp. 520:1–520:8. http://dx.doi.org/10.1145/3290605.3300750.

Schedl, M., 2017. Intelligent user interfaces for social music discovery and exploration of large-scale music repositories. In: Proceedings of the 2017 ACM Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces. HUMANIZE '17, ACM, pp. 7–11. http://dx.doi.org/10.1145/3039677.3039678.

Schrepp, M., Hinderks, A., Thomaschewski, J., 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). Int. J. Interact. Multimed. Artif. Intell. 4 (6), 103–108. http://dx.doi.org/10.9781/ijimai.2017.09.001.

Sen, S., Swoap, A.B., Li, Q., Boatman, B., Dippenaar, I., Gold, R., Ngo, M., Pujol, S., Jackson, B., Hecht, B., 2017. Cartograph: Unlocking spatial visualization through semantic enhancement. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. IUI '17, ACM, pp. 179–190. http://dx.doi.org/10.1145/3025171.3025233.

Senecal, S., Nantel, J., 2004. The influence of online product recommendations on consumers' online choices. J. Retail. 80 (2), 159–169. http://dx.doi.org/10.1016/j.jretai.2004.04.001.

Shneiderman, B., 1992. Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. 11 (1), 92–99. http://dx.doi.org/10.1145/102377.115768.

Shneiderman, B., 1996. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages. IEEE, pp. 336–343. http://dx.doi.org/10.1109/VL.1996.545307.

Sinha, R., Swearingen, K., 2002. The role of transparency in recommender systems. In: CHI '02 Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '02, ACM, pp. 830–831. http://dx.doi.org/10.1145/506443.506619.

Skupin, A., Fabrikant, S.I., 2003. Spatialization methods: A cartographic research agenda for non-geographic information visualization. Cartogr. Geogr. Inf. Sci. 30 (2), 99–119. http://dx.doi.org/10.1559/152304003100011081.

Strauss, A., Corbin, J., 1994. Grounded theory methodology. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research. SAGE Publications, Thousand Oaks, CA, US, pp. 273–285.

Tenenbaum, J.B., Silva, V.d., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323. http://dx.doi.org/10.1126/science.290.5500.2319.

Thorndike, E.L., 1920. A constant error in psychological ratings. J. Appl. Psychol. 4 (1), 25–29. http://dx.doi.org/10.1037/h0071663.

Tintarev, N., Masthoff, J., 2015. Explaining recommendations: Design and evaluation. In: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. Springer US, Boston, MA, pp. 353–382. http://dx.doi.org/10.1007/978-1-4899-7637-6_10.

Tintarev, N., Masthoff, J., 2022. Beyond explaining single item recommendations. In: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. Springer US, New York, NY, pp. 711–756.

Tintarev, N., Rostami, S., Smyth, B., 2018. Knowing the unknown: Visualising consumption blind-spots in recommender systems. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. SAC '18, ACM, pp. 1396–1399. http://dx.doi.org/10.1145/3167132.3167419.

Tkalcic, M., Chen, L., 2015. Personality and recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. Springer US, Boston, MA, pp. 715–739. http://dx.doi.org/10.1007/978-1-4899-7637-6_21.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46 (sup1), 234–240. http://dx.doi.org/10.2307/143141.

Torrens, M., Arcos, J.-l., 2004. Visualizing and exploring personal music libraries. In: Proceedings of the 5th International Conference on Music Information Retrieval. ISMIR '04, pp. 421–424.

Tsai, C.-H., Brusilovsky, P., 2019a. Explaining recommendations in an interactive hybrid social recommender. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19, ACM, pp. 391–396. http://dx.doi.org/10.1145/3301275.3302318.

Tsai, C.-H., Brusilovsky, P., 2019b. Exploring social recommendations with visual diversity-promoting interfaces. ACM Trans. Interact. Intell. Syst. 10 (1), 5:1–5:34. http://dx.doi.org/10.1145/3231465.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Vargas, S., Castells, P., 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11, ACM, pp. 109–116. http://dx.doi.org/10.1145/2043932.2043955.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. 58 (301), 236–244. http://dx.doi.org/10.1080/01621459.1963.10500845.

Xiao, B., Benbasat, I., 2007. E-commerce product recommendation agents: Use, characteristics, and impact. MIS Q. 31 (1), 137–209.

Zhang, Y., Chen, X., 2020. Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr. 14 (1), 1–101. http://dx.doi.org/10.1561/150000006.

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J.R., Zhang, Y.-C., 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. Proc. Natl. Acad. Sci. USA 107 (10), 4511–4515. http://dx.doi.org/10.1073/pnas.1000488107.

Ziemkiewicz, C., Crouser, R.J., Yauilla, A.R., Su, S.L., Ribarsky, W., Chang, R., 2011. How locus of control influences compatibility with visualization style. In: 2011 IEEE Conference on Visual Analytics Science and Technology. VAST, pp. 81–90. http://dx.doi.org/10.1109/VAST.2011.6102445.