Investigating the User Experience of Recommender Systems in Literature and in Movie Streaming Services

Master's Thesis by Patrick Skov Johansen
Supervised by Mikael Brasholt Skov

AALBORG UNIVERSITY
STUDENT REPORT

Interaction Design
Department of Computer Science
Aalborg University

Preface

During the entirety of the master's education in Interaction Design, I have immersed myself in the challenge of designing technologies to augment human abilities. The project work I have participated in, all incorporated an abstraction of artificial intelligence to enable new possibilities or mitigate problematic behaviour. Up until this project, I have worked continuously with two other interaction designers. On my first semester, we proposed a wearable tourist guide concept that matched the user's preferences to that of locals', in the moment. On our second and third semester, we developed and field-tested a self-sorting waste bin. We first studied the clash between citizens and automation technology, and later collaborative artificial intelligence. We wrote a late-breaking work article on that project and got it accepted at the 2020 CHI conference. After passing that project, circumstances drew the three of us our separate ways. As a result of our last project, I was particularly interested in designing the presentation of recommendations from artificial intelligence to make people consider them despite opposing their viewpoint, i.e. how to persuade them. An avenue to conduct such research was recommender systems in movie streaming, where systems and users were extremely accessible. My master thesis did not turn out as I had imagined. I quickly realised that building a recommender system sophisticated enough to evaluate was beyond my abilities, and possibilities to evaluate with users was obstructed by the COVID-19 pandemic. So, I decided to adapt the project and I am satisfied with the outcome.

The Master's thesis I present has two parts: (1) a literature review of recent recommender systems literature focusing on user experience and (2) a report on my empirical investigation of movie streaming user's perception and interaction with recommender systems. My approach was first to gain an overview of the literature to identify an avenue for empirical research (part 1). It became evident that the literature lacked empirical investigations of users' perceptions, behaviour and motives in the context of their everyday lives as opposed to controlled environments. Based on this finding, I interviewed 14 experienced and frequent users of movie streaming services to investigate their user experience and possibly formulate implications for future designers of recommender systems (part 2).

It has been a challenging experience adapting to my own company. I have realised that I am more social than I would care to admit. Being isolated has made me reflect on the power of collaboration. Previously, I could get annoyed if someone did not agree with me, and I would try to convince them. But now I realise that submitting is not a compromise, but an opportunity to improve a shared outcome. I have found it difficult to view my work from another perspective, so I am thankful for having had an inspiring supervisor.

User Experience of Recommender Systems: 2017 and Onwards

Patrick Skov Johansen
Interaction Design
Dept. of Computer Science, AAU
Patrick.skov.johansen@gmail.com

ABSTRACT

User experience research in recommender systems has been gaining momentum in recommender systems literature. To map out recent research efforts concerning the user experience of recommender systems, we conducted a literature review of 53 research articles extracted from the digital library of the Association for Computing Machinery (ACM DL). Our synthesis of reviewed literature was structured in three constructs: Interaction, Algorithm, and Presentation. These constructs constitute the objective system aspects in the user experience evaluation framework proposed by Knijnenburg & Willemsen [27]. Related to each construct, we identified one to three themes. Related to the Interaction construct, we identified the themes: *Preference elicitation & Feedback*, *Input Modality*, and *Interactivity*. Related to the Algorithm construct, we identified the themes: *Diversity*, *Fairness*, and *Context*. Related to the Presentation construct, we identified the theme: *Explanations*.

Author Keywords

Recommender systems; Literature review;

CSS Concepts

Information systems → *Recommender systems*

INTRODUCTION

Recommender systems (RS) have been studied for more than 20 years and continues to draw interest both in research and commercial applications. Traditionally, researchers viewed RS as decision-making tools to alleviate issues with information overload. The early goals of RS were to model the users' preferences as closely as possible and predict their affinity for particular items. Netflix popularised this goal with the Netflix Prize, a $1 million prize to the team that could improve Netflix's recommendation algorithm prediction accuracy by 10%. This area of research is still considerably more active than user experience (UX) research in RS, but an increasing interest has become evident.

In the present work, we have reviewed 56 research articles focusing on UX in RS literature published from 2017 to 2020. In the context of past research efforts, we provide an overview of previous literature reviews RS research focusing on some aspect of user experience. The synthesis of our work is structured according to the objective system aspects in the framework for user experience evaluation in RS by Knijnenburg & Willemsen [27]: Interaction, Algorithm, and Presentation.

BACKGROUND

Recommender algorithms are categorized in three types: (1) Collaborative-filtering, (2) Content-based, and (3) Session-based [18,34]. *Collaborative filtering* recommender algorithms recognise commonalities between users based on their interaction with items and generate recommendations based on the comparison of these users. It is based on the assumption that people who like the same items will continue to do so. Item-item and user-item collaborative filtering algorithms are variations of this. *Content-based* recommender algorithms learn the users' preferences based on the features of the items they interact with. Then it recommends items that have similar features. *Session-based* recommender algorithms do not rely on users' interaction history but on in-session queries alleviating a common problem called "the cold-start problem" where the features of new users or items are unknown in the beginning.

In the following section, we will summarise previous reviews in RS literature focusing on user experience. We identified seven literature reviews, each focusing on different subtopics related to the user experience of RS. The subtopics range from recommendations methods, human decision making, and the evolution of recommender systems, to the interaction aspects of recommender systems. The focus of this review is to outline the developments in recent RS. Our work contributes knowledge of where research efforts have focused and identify which areas have received less attention.

Reviews of Recommender Systems literature

Of the reviewed surveys, see Table 1, the paper by Adomavicius & Tuzhilin [2] is positioned closest to the traditional approach aimed at optimising the algorithms themselves. The reason this review was included, was that they presented early challenges with recommender systems, such as "New User Problem" and "New Item Problem". These problems continue to be a challenge in RS research under the collective title "the cold-start problem". The cold-start problem has more recently become a focus area for HCI researchers who experiment with different ways of eliciting user preferences without burdening the users [23]. More importantly, Adomavicius & Tuzhilin [2] discuss the possibility of extending RS with contextual information [1] that may be crucial in some applications where, for example, the utility of a particular product may depend significantly on the temporal context of the user.

Author(s)	Year	Focus	Papers surveyed or [references]	Publication
Adomavicius & Tuzhilin [2]	2005	Recommendation methods/techniques	[112]	IEEE Transactions on Knowledge and Data Engineering
Konstan & Riedl [28]	2012	Evolutions of Recommender Systems	[93]	User Modeling and User-Adapted Interaction - The Journal of Personalization Research
Chen, De Gemmis, Felfernig, Lops, Ricci & Semeraro [7]	2013	Human decision making & Recommender Systems	[21]	ACM Transactions on Interactive Intelligent Systems
Amatriain & Basilico [3]	2016	Recommender systems in the industry	[32]	RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems
He, Parra & Verbert [18]	2016	Interactive recommender systems	[125]	Expert Systems with Applications - An International Journal
Valdez, Ziefle & Verbert [52]	2016	HCI terms in the recommender systems literature	9.432	RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems
Jugovac & Jannach [23]	2017	User interaction aspects in recommender systems literature	[216]	ACM Transactions on Interactive Intelligent Systems

Table 1: Overview of reviews included in the summary of UX research in recommender systems literature. Numbers in [x] are the number of references for the given review if they do not report how many papers they surveyed.

Seven years later, in 2012, Konstan & Riedl [28:108] contributed a historical overview of the evolution of recommender systems titled: "From Algorithms to User Experience". In their review, they defined user experience in RS as:

"By user experience we mean the delivery of recommendations to the user and the interaction of the user with those recommendations" [28:108]

Compared to Jugovac & Jannach [23], who presents a framework which considers both preference elicitation and recommendation presentation and interaction, it is limited to only consider the interaction with the presented recommendations. Nevertheless, Konstan & Joseph [28] report four themes from their review concerning: (1) The user-recommender lifecycle, (2) notions of quality beyond prediction

accuracy, (3) risks of recommenders, and (4) giving users more control over recommendations. Within these, they touch upon topics such as contextual information, recommendations for groups, transparency, interactive recommendation dialogues, and the cold start problem, to name a few. Lastly, they point to a challenge within the research community of conducting more user studies, especially field studies.

Chen et al. [7] contributed a less comprehensive study, however, differentiated by drawing from cognition theory related to human decision-making. They conclude with open issues on group decision; much research focuses on single-user scenarios, the impact of contextual factors; defined as "*any information or conditions that can influence the perception of the usefulness of an item for a user*" [7:5], and lastly personal factors; such as personality, mood or emotions.

Amatriain & Basilico [3], from Quora and Netflix respectively, contributed a review of RS in the industry. In their review, they provided examples of how RS was being used across domains and argued that each domain had to deal with unique challenges. They explained that recommendations have become ubiquitous and with such an emphasis on recommendations, an important element is to provide awareness and explanations for why an item was recommended. RS in the industry relied mostly on implicit feedback, contradictory to most research. They explained that this approach results in only positive and missing data and no negative data. Additionally, the interpretation of implicit data can be misleading. They propose research efforts in using indirect feedback that includes contextual information to alleviate this issue partly. Finally, they propose that we should investigate personalising not only what we recommend but how we recommend.

He et al. [18] contributed an "interactive visualisation framework" for recommender systems, which they used to analyse 24 interactive recommender systems that have been proposed in the research community. They focus especially on interactive visualisations, e.g. node-link diagrams and

#	Query: Research Article from 2017 - 2020	Results	Results Surveyed	Papers retrieved	Overlap
1	Recommend	21.611	200 (0,9%)	21 (10,5 %)	-
2	"Recommender system?" and "user experience"	253	253 (100%)	41 (16,2 %)	8 (2,3 %)
3	"Recommender system?" and "user experience" and keywords: "Recommend"	67	67 (100%)	3 (4,5 %)	23 (23,7 %)
4	CHI: "Recommender systems"	68	68 (100%)	7 (10,3 %)v	0 (0 %)
Total papers retrieved				72	
Total papers eligible for this review				**53**	

Table 2: The search strategy used for this review was an initial broad search followed by incrementally more specific queries. The results were sorted based on "relevance" on ACM DL.

scatterplots. They argued that the purpose of visualising recommendations is to achieve the following objectives: transparency, justification, controllability, diversity, mitigate the cold start problem, or to incorporate contextual information. They present transparency and justification as distinct forms of explanations. Transparency deals with explaining the "black-box" / inner workings of the algorithm. Justification aims to help the user understand why they get certain recommendations, but it may not relate to the inner logic of the recommendation techniques. They also discussed "serendipitous recommendations" meant to help users discover new and interesting items that they might not have discovered otherwise. They argued that interactive visualisations techniques support users in exploration and discovering "non-obvious" recommendations.

Valdez et al. [52] contributed a bibliometric analysis of RS literature from the Corpus database. Their reports primarily summarise the work of He et al. [18], even though much more work exists. Nevertheless, they conclude that HCI related aspects are underrepresented in RS literature and propose four future research topics to investigate: user control, adaptive, affective, and high-risk domains.

Lastly, Jugovac & Jannach [23] provided a comprehensive review of interaction aspects related to RS. In their work, they proposed a framework separating preference elicitation and result presentation. The framework listed sub-elements under each, which they reviewed and discussed separately throughout. Following the presentation of results of each section, is a discussion of emergent research challenges. Related to preference elicitation, they present five areas: (1) *Biases*; feedback mechanism that avoids bias in collected data and helping users state their preferences more consistently. (2) *Detailed feedback*; more fine-grained feedback. (3) *User engagement*; stimulating the user to give more feedback and recognising user interest drifts. (4) *Novel interaction methods*; e.g. natural language in dialogue-based systems. (5) *Adaptations*; tailoring elicitation approach to the user's needs, determining the next interactional move in conversational systems, and combining long-term preferences with short-term needs. Related to recommendation presentation and user feedback they present six areas: (1) *List design*; determining the optimal choice set size for a given user and application domain, organising interfaces with multiple lists, avoiding boredom and creating diversified lists. (2) *Visualisation*; helping users understand the relationships between items (and other users) through interactive visualisations and designing easy-to-comprehend visualisations approaches that can be integrated into real-world systems. (3) *Explanations*; explaining the differences between choices to the users and generating interpretable persuasive explanations from complex machine-learning models. (4) *User control*; allowing users to give feedback on the recommendations in an intuitive way. (5) *Timing*; deciding when to recommend in proactive recommendation scenarios. (6) *Methodology*; developing standardised evaluation methodologies for novel interaction mechanisms.

To summarise; much has happened since 2005, and it is evident that user experience has received increasing interest within the RS literature. With this summary of reviews focusing user experience of RS, we see recurring themes around contextual information, transparency, user control, explanations and acknowledgement of serendipitous recommendations to avoid boredom.

The User Experience of Recommender Systems

Knijnenburg & Willemsen [27] presented a user-centric evaluation framework for recommender systems. They defined user experience (UX) as *"user's self-relevant evaluations of the qualities of the recommender system"*. The framework consists of six interrelated concepts: (1) Objective Systems Aspects, (2) Subjective System Aspects, (3) User Experience, (4) Interaction, (5) Personal Characteristics, and (6) Situational Characteristics. Each concept consists of a lower-level classification of recommender systems related constructs, see Table 3. We will briefly present Objective System Aspects, User Experience, and Interaction in the following.

The framework emphasises that each concept affects user experience and interaction with the recommender system. The objective systems aspect constructs; Interaction, Algorithm, and Presentation, affect the user experience but is mediated through the subjective systems aspects–the user's perception of the system; usability, quality, and appeal.

The user experience concept relates to three recommender systems constructs. (1) evaluation of the recommender system itself (e.g. perceived system effectiveness). (2) evaluation of the process of using the system (e.g. expressing preferences and choosing recommended items). (3) evaluation of the chosen items (e.g. choice satisfaction).

The last step is the interaction with the system. The interaction is the observable behaviour in which the subjective aspects are grounded. While the subjective aspects, provide the reasoning for the interaction.

Concept	Constructs
Objective System Aspects	Interaction, Algorithm, Presentation
Subjective System Aspects	Usability, Quality, Appeal
Personal Characteristics	Gender, Privacy, Expertise
Situational Characteristics	Routine, Trust, Goal
User Experience	System, Process, Outcome
Interaction	Rating, Consumption, Retention

Table 3: Top-level concepts and lower-level constructs in the user-centric evaluation framework of recommender systems proposed by Knijnenburg & Willemsen [27].

METHOD

In this paper, we report a literature review on research articles from 2017 - 2020 (May) focusing on user experience in recommender systems. Our approach to this literature review was inspired by the method presented by Kitchenham [25]. Based on this method, the process commenced in four steps: (1) Identification & Screening, (2) Examination & Data Extraction, (3) Exclusion Survey, and (4) Synthesising results. 53 papers were eligible for this literature review.

Research Question

This literature review was carried out to clarify what research efforts has been made regarding user experience of RS and investigate in which areas the community has brought the baton from past research. Our research questions for this work was:

1. *What is the focus of recent research focusing on the user experience of recommender systems, and what are the limitations and opportunities for future research?*

Research Strategy

To identify primary studies, we searched ACM Digital Library (DL), inspired by [41]. To identify past literature reviews, presented in the Background section, we searched both ACM DL and Google Scholar.

The search strategy for primary studies was to start broadly and incrementally narrow the scope with more specific queries. Lastly, we searched specifically in the CHI conference because it seemed underrepresented. See Table 2 for the results of the search strategy.

Selection Criteria & Process

To decide what research articles to include we used four selection criteria: (1) research articles must have been published since 2017, (2) research articles must explicitly position itself within recommender systems – articles that investigate systems with an RS component but choose not to position it as such explicitly (through keywords, title, abstract) was not included in this review, (3) research articles must include a study, and (4) research articles must focus on understanding or improving the user experience of RS.

The process of selecting research articles for this review was carried out in three steps: Identification & Screening, Examination & Data Extraction, and lastly an Exclusion Survey.

Identification & Screening

Queries to identify studies was carried out using the search engine on ACM DL. Possibly relevant research articles were first identified based on the title. After that, we screened the abstract on the detail-page and assessed it on relevance to the research question and the selection criteria. Seventy-two research articles passed the screening.

By reading abstracts, it became evident that many papers on ACM DL focus on algorithmic optimisation. Such a focus excludes it from this review, unless if it focuses on improving UX. If these research articles did not mention or hint at a focus on user experience in the abstract or title, it would have been excluded before full-text retrieval. An example of this is [8] with the title: "Learning to recommend accurate and diverse item". The title indicates a relation to serendipitous recommendations mentioned in the Background section. However, the abstract revealed that it concerned algorithmic optimisation. Table 2

Examination & Data Extraction

The examination commenced by reading the retrieved research articles. We focused on the introduction, method, and discussion. Using this strategy helped to identify the focus and contributions of the work and the authors' reflections on it concerning the field of RS.

During the examination, we extracted seven types of data from each research article (if available): research question, subject of study, application domain, individual or group subjects, evaluation methodology, number of participants, and keywords formulated by us. We extracted the data into a spreadsheet, where we could add notes to each entry.

Exclusion Survey

After the examination of the research articles was complete, we surveyed each entry in the spreadsheet to evaluate which articles to exclude. We judged each entry by the selection criteria resulting in nineteen articles being excluded:

- 4 for not meeting selection criteria 2; *Research articles must explicitly position itself within RS.*
- 13 for not meeting selection criteria 3; *Research articles must include a study.*
- 2 for not meeting selection criteria 4; *Research articles must focus on understanding or improving the user experience of RS.*

RESULTS

A user's interaction with RS can be simplified to three overall steps. (1) The user provides information about their preferences; explicitly or implicitly. (2) The algorithm processes the information and generates recommendations predicted to maximise some metric (e.g. choice satisfaction). And (3) the presentation of the recommendations. Then, the user-recommender interaction proceeds iteratively with continuous recommendations and preference input. This interaction is the recommender system experience, and the user's perception evaluation is what constitutes the user experience of recommender systems according to Knijnenburg & Willemsen. [24]. The concept of *objective system aspects*, in the framework proposed by Knijnenburg & Willemsen [27], encapsulate these steps in the three constructs: *Interaction (Input)*, *Algorithm (Processing)*, and *Presentation (Output).*

Knijnenburg & Willemsen [27]	**This review**
Construct	**Topics**
Interaction	Preference elicitation & Feedback
	Input Modality
	Interactivity
Algorithm	Diversity
	Fairness
	Context
Presentation	Explanations

Table 4: The concept Objective System Aspects from Knijnenburg & Willemsen [27] consists of three constructs. Related to each construct are a number of topics identified from synthesising the literature in this review.

Interaction (Input)

Knijnenburg & Willemsen [27] defines the construct *interaction* as the input mechanisms the user interface with to express their preferences and communicate with the recommender system in general [27]. Our synthesis of reviewed literature identifies three topics that investigate the objective interaction mechanisms of recommender systems: preference elicitation & feedback, input modality, and interactivity. See Table 4.

Preference Elicitation & Feedback

Preference elicitation in RS concerns aspects related to extracting accurate user preferences to create the foundation for personalised recommendations. These preferences are extracted either through explicit feedback (e.g. the user rates an item) [20,42], or implicit feedback (e.g. the user hovers on a recommendation) [13,44,53]. Preference elicitation is often employed when onboarding new users in a system (e.g. YouTube [9]) to overcome the cold-start problem [49]. There is an ongoing discussion in the literature of what type of feedback to rely on to extract accurate data about a user's preferences [57]. Explicit feedback is found to be very accurate but can increase the cognitive load of the user [16,22]. Implicit feedback is very accessible, but it is difficult to interpret [3].

Schnabel et al. [44] reported on the interpretability of different implicit feedback. They conducted a study of users' tolerance for receiving less accurate recommendations intended to explore the user's preferences, and exploration strategies to mitigate the cost. They found that limiting the amount of exploration, to a certain threshold, was essential both for user satisfaction and implicit feedback quantity such as hovering and short-listing. Additionally, they found short-listing to be of higher quality than hovering which, as they report, *"support to the insight that learning algorithms can greatly benefit from improved interface design where users are offered the right incentives to interact"* [44:520].

Novel approaches to implicit feedback have also spurred in recent years. Studies of both *negative experiences* [33] and *inaction* [58] as implicit feedback modes have been conducted. Lu et al. [33] studied the effects of negative experiences in news streaming and automatic detection of it. Unsurprisingly they found that negative experiences affect user satisfaction in the current session, but also found that the impact carries over to the next session. Zhao et al. [58] conducted a field survey in a live movie recommender system to interpret what inaction means from the perspective of the user and the system. They propose seven categories of reasons for user inaction that they predict from log data. They argue that not all inaction should be treated as negative feedback and inferring the reason for inaction is essential for increasing user action engagement.

Zhao et al. [57] conducted a 1-month investigation reported in the article: "Explicit or implicit feedback? Engagement or Satisfaction?". They compare six different algorithms and find that the ones optimising for implicit action error is more engaging but are less accurate. They advise using a hybrid of implicit and explicit feedback to mitigate the costs regarding accuracy. They conclude that the selection of algorithm in machine learning recommender systems significantly affects the user experience, as reflected in the framework by Knijnenburg & Willemsen. [27].

Input Modality

Input modality regards the channels of input from the user to the system. In the latest review by Jugovac & Jannach [23], they reported on emergent research challenges concerning novel interaction methods, primarily pointing towards dialogue-based natural language interfaces. Kang et al. [24]

studied peoples' first-time use of a natural language interface for movie recommendations. They found that people make follow-up queries to correct their initial query in a critiquing style while many others reformulate or start over completely. Additionally, they compared speech and typing inputs. They found speech queries to be both longer and contain richer details about the user's preferences. Other studies have utilised natural language recognition for their input modality, such as Chatbots [56], but do not study the effects of the modality on user experience.

Interactivity

Interactivity integrates user-control to mitigate issues with algorithmic blind-spots, transparency [16], and improved perception of recommendations [21]. Increased user-control is correlated with more reliance on explicit feedback, as opposed to implicit feedback, regarding different aspects of the experience. The research on this topic either allow the user to control the weights of different situational or preference aspects [16,21,22] or display visualisations to enable user-initiated discovery to increase perceived diversity [15,50].

Harambam et al. [16] conducted focus groups (N=21) to study people's perception of different user control mechanisms presented as mockups in the context of interactivity. Their findings show that their participants generally desire more control to mitigate their perceived loss of agency due to RS. Additionally, the participants were sceptical of what control the mechanisms offered, which might be due to the static nature of their study. Jin et al. [21] evaluated user control in a music recommendation system on Amazon Mechanical Turk. Participants were able to control musical preference weights and contextual factors. They found that affording control lead to higher perceived quality with no increase in cognitive load, contrary to other studies in the literature [16,22]. They also found that people adjust their weights during relaxing times, suggesting that context affects interactivity as well. Tsai et al. [50] evaluated three RS interfaces for social recommendations at scientific conferences. Their interfaces were designed to explore user-control and diversity-awareness. Their findings show that their user-control enhanced interfaces directly affect the system's perceived diversity.

Conversational interfaces are a category related to interactivity. These interfaces are turn-based as opposed to traditional direct manipulation interfaces. The interaction unfolds with the user initiating an interaction, where the system asks for input, and the user supplies information in an iterative "conversation" until the system can present a recommendation. Conversational RS is more engaging, improve preference elicitation and offer better recommendations [9]. Enabling this interactivity gives the user a perception of control and increases the perception of transparency [40]. Conversational interfaces in RS have been investigated in the forms of chatbots [20,56], user onboarding flows [9], and desktop GUI applications [49]. In these studies, the effects were measured in terms of choice satisfaction and engagement.

In summary of the interaction construct, many researchers have investigated preference elicitation and feedback mechanisms to enable it. There are discussions of whether to rely on explicit or implicit feedback and so far, a middle ground is advised to compensate for engagement and accuracy. Novel implicit feedback modes have also been investigated. Negative experiences and inaction can both contribute to better preference elicitation. Negative experiences can, when detected, be used as feedback to mitigate the intra-session effect of the experience. Inferring the reasons for user inaction can be used to increase user action engagement with the recommended items. Investigation of using speech as input modality has also been conducted. Findings show that using speech for input is promising, at least in movie RS, because participants give richer details about their preferences compared to typing. Evaluations of interactivity have also been conducted. This topic of research has aimed at increasing user-control or improving perceptions of the recommendations, e.g. diversity.

Algorithm (Processing)

The *algorithm* construct is defined as the processing of information to deliver recommendations [27]. The synthesis of reviewed literature identifies three topics about incorporating different parameters in the algorithmic processing to enhance the user experience: diversity, fairness, and context.

Diversity

Steck [42] defines diversity as "minimal redundancy or similarity among recommended items"[46]. It is the countermeasure to the over-personalisation that comes with accuracy optimisation. When algorithms are overly optimised for accuracy, they create, so-called, "filter-bubbles" [36,55] that negatively affects user satisfaction [55] and leave the users bored [35]. Diversified recommendations are naturally inaccurate but contribute to better item discoverability and satisfaction [13,50]. Zanitti et al. [55] present a diversity by design recommender system in the movie application domain. Their proposal incorporates enhancement on four dimensions of diversity: global coverage, local coverage, novelty, and redundancy. They evaluated their proposed algorithm on the Movielens dataset where it performs comparably to state-of-the-art recommenders.

Serendipity is a concept closely related to diversity in RS literature and has recently been studied considerably [13,31,35,39,54]. Serendipity-oriented methods are expected to recommend unfamiliar surprising items beyond the user's discovery [32], thus countering over-personalisation. Li et al. [32] proposed DESR, the Directional Explainable Serendipity Recommendation method. Their contribution is predicting serendipitous recommendations with regards to users' long-term preferences and short-term demands. They evaluated DESR on the Movielens dataset, and results show significant improvement compared to state-of-the-art serendipity recommender systems. Maccatrozzo et al. [35] investigated users' coping potential for new items, e.g. ability to appreciate serendipitous recommendations. They conducted an online experiment with TV-programme recommendations and validated their assumptions about people with high coping potential are more inclined to accept serendipitous recommendations.

Discovery is yet another concept related to diversity in RS literature. Garcia-Gathright et al. [11] defined discovery as the experience of finding content that is previously unknown to the user. It emphasises the active role of the user compared to serendipity and diversity. Garcia-Gathright et al. [13] studied the use and evaluation of a system for supporting music discovery. Through their mixed-methods approach, they identified four user goals that influence their behaviour. Users engaging in discovery expect it to be hit-and-miss, i.e. one satisfactory item is enough to create a positive experience. They also predict user satisfaction from logged interactions and confirm that inferring user goals, along with capturing interactions per item, contributed significantly to this achievement.

Any recommendation algorithm faces a trade-off between exploiting incomplete knowledge of the user's preferences to maximise satisfaction in the short term and *discovering* additional user preferences to maximise satisfaction in the long term [36,44]. The recommendation algorithm elicits feedback on items of uncertainty to learn more about the user's preferences, called *algorithmic exploration* [44]. Algorithmic exploration is similar to *discovery* but is initiated by the algorithm without the user’s explicit knowledge. Schnabel et al. [44] studied people’s tolerance for algorithmic exploration in six different levels of exploration. Their findings showed no significant effect of exploration on overall satisfaction, which implies that people find ways to be successful at their task despite the decreased accuracy. McInerney et al. [36] presented BART, BAndits for Recsplanations as Treatments, a contextual bandits-based framework that addresses the problem of recommending with explanations under uncertainty of user satisfaction. The purpose of BART is to learn how items and explanations interact within any given context to predict user satisfaction. In both offline and online evaluations, their results show that people respond differently to explanations and the recommendations outperform static ordering of explanations by 20%.

Fairness

In recent RS literature, fairness has not employed consistently. Authors have adapted the definition of fairness to their problem domain, such as Hezog & Wörndl [19] who study recommendation for groups. Their definition is equal representation of an individual's preferences in the recommendation. Recommendations for groups is the most prominent area of research with regards to fairness [19,45,48]. The goal of optimising fairness is to satisfy each individual in the group. The discussion is on how to approach the problem of modelling preferences in a group constellation; whether to aggregate individual preferences [19,45,48] or model the unique interaction when in a group [5]. Cao et al. [5] presented AGREE, Attentive Group REcommEndation, that aims to learn the influence of a member, and how to adapt the influence when the group interacts with different items. They evaluated on two datasets from tourism and movies. Results show that AGREE performs significantly better than state-of-the-art methods.

Steck [46] proposes a complementary definition of fairness that recommendations should reflect the user’s various interest to their corresponding proportions. Their approach is termed “Calibrated Recommendations” and aims to optimise for diversified recommendations in the full spectrum of the users stated preferences. The algorithm's goal is to overcome a common problem in machine learning recommender systems where a user's main interests are overrepresented compared to lesser interests.

Context

Context-Aware Recommender Systems (CARS) is recommender systems capable of factoring situational characteristics when making recommendations. CARS have been an emergent topic for some time [6,21,36,37,58]. Jin et al. [21] evaluated a prototype system giving the user explicit control of the contextual input variables: weather, social, time and mood. They find that having control of these variables leads to higher perceived quality and no additional cognitive load. Mei et al. [37] proposed a neural model that adaptively captures the interactions between contexts and user-item. They verify that contexts have a significant influence on people's preferences and results show that their proposed model outperforms the state-of-the-art on both rating prediction and personalised rankings tasks.

Zheng [59] presented an inverse utilisation of context. Instead of recommending items based on the present context, recommend contexts for particular items. They investigated whether user's preferences on context could be inferred by contextual ratings and compared indirect context suggestion to direct context predictions. They found indirect context suggestions to perform better.

In summary of the algorithm construct, a focus on diversifying recommendations has emerged and with that, a shift from focusing on providing the *most* accurate recommendations to recommendations that give the most satisfaction. This research is complemented by efforts on discovery; supporting users in developing their preferences, and algorithmic exploration; eliciting feedback on items with uncertain user satisfaction. Fairness has been investigated in the context of group recommendations to satisfy each member of the group. Issues of how to aggregate user pretences in group constellations have been investigated, and two approaches seem to be prevailing. Context has been verified to influence people's preferences significantly, and efforts of factoring the item-context and user-context relationships when generating recommendations have been reported.

Presentation (Output)

Knijnenburg & Willemsen [27] defines the presentation construct as the presentation of recommendations to the user. Visual design considerations, such as the number of recommendations to present and their layout, are naturally a part of the presentation. However, since we did not identify articles focusing on the effects of such aspects on user experience, this will not be presented. Our synthesis of reviewed literature identifies one topic about presenting recommendations to enhance the user experience of RS: explanations.

Explanations

There are two types of explanations in RS literature: "white-box" and "black-box" [40]. White-box explanations convey input, output, and the steps to arrive at a decision. They focus mainly on the system's reasoning and fill a gap between the user's perception and the system's internal process. Black-box explanations provide justifications for the system and its outcomes without disclosing how the system works. This distinction between explanations is similar to that of He et al. [18] presented in the Background section.

Explanations in recent RS literature have been employed to increase trust [4,26,30,32], increase user engagement [36,38], increase awareness of how a system works [38,40], and justify why particular recommendations are made [29,56]. Kunkel et al. [30] investigated the effects of explanations delivered in natural language (from people) versus state-of-the-art RS (algorithmic output) on trust in the recommendation source. They find that a useful explanation can, to a certain degree, make up for an inaccurate recommendation. They explained that there are intrinsic social components in the way people deliver recommendations that current RS cannot compete with. They concluded that the tremendous accuracy in automated RS is next to meritless when it fails to convey the rationale behind its recommendations. Kleinerman et al. [26] investigated explanation in reciprocal environments, i.e. why both parties are expected to benefit from the match in, e.g. dating or job recruitment. They conducted empirical evaluations with 287 human participants using a dating app. Their results showed that reciprocal explanation outperforms other explanation methods when a high monetary or emotional cost is involved. However, when this cost is not present, reciprocal explanations were found to perform worse than other methods. Kouki et al. [29] studied the problem of generating personalised explanations for hybrid recommender systems with many data sources. They conduct a crowd-sourced study with real users on the last.fm music platform. They found that users prefer item-centric explanations as compared to user- or socio-centric explanations. Participants in their study also rated textual explanations as more persuasive than visualisations.

Explanations that aim to increase users' awareness of how the system works, i.e. transparency, has been studied in various application domains, allowing the users to create fitting perceptions of e.g. controllability, and scrutability [11,40]. Optimising transparency is often advised as a means to increase trust, but increasing transparency sometimes comes with trade-offs. Gómez-zará & Dechurch [14], studied the effects of displaying diversity in a recommender system for self-assembling teams. They found that users avoid team members who increase diversity in their team and instead select team members who are similar, contrary to social norms. Rader et al. [40] studied how five different explanation styles affected their participant's perception of Facebook's News Feed algorithm. They report that all styles made participants more aware of how the system works, which helped determine whether the system is biased and if they can control its output. They discuss that accountability is the ultimate goal of transparency mechanisms via increased scrutiny. However, with systems that have a higher degree of agency than the user, transparency is disconnected from power.

In this section, the results of reviewing user experience focused on RS literature from 2017 to 2020 has been expounded. Many investigations have been conducted, and there is still interest in improving the user experience of RS. The majority of the reviewed literature has been conducted with regards to input; how users communicate their preferences to the system and allowing users to explore the item repository on their terms. Still, we have laid out all objective system aspects of the RS user experience.

DISCUSSION

Recommender systems have been a topic of research for more than 20 years. In the early days of recommender systems research, the focus was to model the preferences of the user as accurately as possible and mitigate choice overload. More recently, efforts focusing on the user experience of recommender systems have been conducted. In the present work, we have presented our synthesis of 53 research articles focusing on improving the user experience of RS published from 2017 to May 2020. We argue that these efforts are guided by previous research conducted in the field, why we have also presented an overview of previous reviews and the opportunities for future research they identified. We structured our synthesis according to the objective system aspects in Knijnenburg & Willemsen's [27] framework for user experience evaluation in RS. Our findings show that research related to all the objective system aspects proposed by Knijnenburg & Willemsen's [27] have been conducted. Compared to the other aspects, research efforts concerning Presentation aspects of RS are less varied, focusing predominantly on explanations. A possible reason for this is that presentation of information is not a research topic exclusive to RS, and that the related research is more present in, e.g. information visualisation literature. Nevertheless, explanations have been a popular topic that attracted much interest, especially with the emergence of black-box machine learning in RS and beyond.

Surprisingly few research articles [11,21,22,43,51,54,58] in our review employ the framework and guidelines by Knijnenburg & Willemsen [27]. Possible explanations for this can be the variety of research disciplines included in this review. Studies that mainly focuses on algorithmic optimization evaluate their inventions in either offline environments; processing public datasets and simulations studies, or online; in live products with actual users or customers without them knowing. These studies do not rely on questionnaires to measure the constructs devised in the framework by Knijnenburg & Willemsen [27]. Instead, they interpret implicit user feedback and attempt to predict user satisfaction; one of the constructs of user experience in the framework.

We identified user satisfaction, or choice satisfaction, as the most used metrics to evaluate improvements in the user experience of RS. Other parameters also exist, such as system effectiveness and choice difficulty [27]. Knijnenburg & Willemsen [27] argued that it was essential also to consider the subjective system aspects to explain changes in experience and interaction as a result of a particular treatment, e.g. [22]. Schaffer et al. [43] also identified the reliance on measuring satisfaction as the only indication for user experience. They similarly suggested the importance of investigating why these effects came about and proposed a new methodology to measure choice satisfaction independently of user experience.

When we examined how user experience has been studied in recent RS literature, it was evident that most studies failed to define what they mean by the term or whose definition they rely on. Our impression is that the general understanding is to optimise user experience, and that satisfaction is a good way to measure it. Knijnenburg & Willemsen [27] defined the user experience of RS as the *"user's self-relevant evaluations of the qualities of the recommender system"*. In their framework, the evaluation is of the system, the process of using it, and the user's satisfaction with their choice. They emphasise that these evaluations are based on the user's perception of the objective system aspects. Their framework also demonstrated that personal characteristics and situational aspects affect the user experience. However, as compared to Hassenzahl [17], there is no consideration of the user's driver for the experience, i.e. what the satisfaction is meant to fulfil. Hassenzahl defined user experience as a *"momentary, primarily evaluative feeling (good-bad) while interacting with a product or service."*, which is comparable with Knijnenburg & Willemsen's [27] definition with a few exceptions. Hassenzahl [17] extended his definitions with: *"Good UX is the consequence of fulfilling the human needs for autonomy, competency, stimulation (self-oriented), relatedness, and popularity (others-oriented) through interacting with the product or service (i.e., hedonic quality). Pragmatic quality facilitates the potential fulfilment of be-goals."*

Contrary to Knijnenburg & Willemsen, Hassenzahl did not separate perception and experience evaluation but considered both as constituents of user experience. Many of the constructs in the framework by Knijnenburg & Willemsen can be regarded as contributors to a product's pragmatic quality, and their framework did not include consideration for the psychosocial needs associated with a user's interaction with a system. We argue that this aspect has been overlooked and that the user's motives for using an interactive product are essential to evaluate the fulfilment of psychosocial needs and in extension, the user experience. Knijnenburg & Willemsen's [27] framework does emphasise that situational characteristics affect the user experience. This observation only provides support that understanding of the psychological needs is necessary because a user's motivations for using an interactive product are momentary and can change from session to session.

Few articles include considerations for a user's mood [13,35,36,44], which is similarly situational and found to be affected by the time of day and day of the week [12], daily events [47], and the weather [10]. A user's mood might be related to the situational psychological needs of the participants, but none of the reported studies investigated the effects of this. In general, the articles in our review could be improved with a more user-centred approach to understanding the user experience of RS. This includes using appropriate methods to elicit this information and evaluating systems in a context that induces similar to real-life needs for the user. Many of the studies were evaluated in controlled environments detached from the user's everyday lives, e.g. on crowdsourcing platforms or evaluating experimental features online.

Moving forward, we suggest that future studies in user experience of RS considers the following three suggestions. (1) Identifying application domain-specific challenges and address them with the appropriate means. For example, many studies in our review measured trust in various application domains. We argue that trust may be more important in high-risk application domains than in low-risk application domains. (2) Be specific about the theories used. The majority of the 53 papers in our review, did not mention what user experience theory they relied on or used an established user experience evaluation framework. (3) Apply user-centred methods to gain a better understanding of the user's motives. What the user intents to accomplish by using an interactive product, is according to the user experience definition by Hassenzahl [17], the very foundation for the user experience.

CONCLUSION

In this article, we have reported our review of 53 research articles focusing on improving the user experience of recommender systems published from 2017 to May 2020. Our work revealed a gap in how the RS literature operationalises the term user experience, compared to HCI literature. The majority of the reviewed literature failed to define their understanding of user experience or whose definition they rely on. Satisfaction is the primary metric for evaluation with no regard to what needs the user seeks to fulfil. We suggest that future researchers focusing on UX in RS considers the appropriate means for the application domain, are specific about the theories they use, and apply more user-centred methods, rooting their research in the motives of the users or interactive technology.

REFERENCES

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems* 23, 1 (January 2005), 103–145. DOI:https://doi.org/10.1145/1055709.1055714

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering 17*, 734–749. DOI:https://doi.org/10.1109/TKDE.2005.99

[3] Xavier Amatriain and Justin Basilico. 2016. Past, present, and future of recommender systems: An industry perspective. In *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 211–214. DOI:https://doi.org/10.1145/2959100.2959144

[4] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to recommend? User trust factors in movie recommender systems. In *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, New York, USA, 287–300. DOI:https://doi.org/10.1145/3025171.3025209

[5] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. 2018. Attentive group recommendation. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Association for Computing Machinery, Inc, New York, NY, USA, 645–654. DOI:https://doi.org/10.1145/3209978.3209998

[6] María del Carmen Rodríguez-Hernández, Sergio Ilarri, Raquel Trillo, and Ramon Hermoso. 2017. Context-Aware recommendations using mobile P2P. In *MoMM2017: Proceedings of the 15th International Conference on Advances in Mobile Computing & Multimedia*, Association for Computing Machinery, New York, New York, USA, 82–91. DOI:https://doi.org/10.1145/3151848.3151856

[7] Li Chen, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro. 2013. Human decision making and recommender systems. *ACM Transactions on Interactive Intelligent Systems* 3, 3 (October 2013), 1–7. DOI:https://doi.org/10.1145/2533670.2533675

[8] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In *26th International World Wide Web Conference, WWW 2017*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 183–192. DOI:https://doi.org/10.1145/3038912.3052585

[9] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. 2018. Q&R: A two-stage approach toward interactive recommendation. In *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY, USA, 139–147. DOI:https://doi.org/10.1145/3219819.3219894

[10] Jaap J.A. Denissen, Ligaya Butalid, Lars Penke, and Marcel A.G. van Aken. 2008. The Effects of Weather on Daily Mood: A Multilevel Approach. *Emotion* 8, 5 (October 2008), 662–667. DOI:https://doi.org/10.1037/a0013497

[11] Vicente Dominguez, Ivania Donoso-Guzmán, Pablo Messina, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, NY, USA, 408–416. DOI:https://doi.org/10.1145/3301275.3302274

[12] Boris Egloff, Anja Tausch, Carl Walter Kohlmann, and Heinz Walter Krohne. 1995. Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion* 19, 2 (June 1995), 99–110. DOI:https://doi.org/10.1007/BF02250565

[13] Jean Garcia-Gathright, Brian St Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Association for Computing Machinery, Inc, New York, NY, USA, 55–64. DOI:https://doi.org/10.1145/3209978.3210049

[14] Diego Gómez-zará and Leslie A Dechurch. 2020. The Impact of Displaying Diversity Information on the Formation of Self-assembling Teams. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. DOI:https://doi.org/https://doi-org.zorac.aub.aau.dk/10.1145/3313831.3376654

[15] Francisco Gutiérrez, Sven Charleer, Robin de Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. 2019. Explaining and exploring job recommendations: A user-driven approach for interacting with knowledge-based job recommender systems. In *RecSys 2019 - 13th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 60–68. DOI:https://doi.org/10.1145/3298689.3347001

[16] Jaron Harambam, Mykola Makhortykh, Dimitrios Bountouridis, 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 *RecSys 2019 - 13th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 69–77. DOI:https://doi.org/10.1145/3298689.3347014

[17] Marc Hassenzahl. 2008. User experience (UX). In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08*, ACM Press, New York, New York, USA, 11. DOI:https://doi.org/10.1145/1512714.1512717

[18] 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. DOI:https://doi.org/10.1016/j.eswa.2016.02.013

[19] Daniel Herzog and Wolfgang Wörndl. 2019. A user study on groups interacting with tourist trip recommender systems in public spaces. In *UMAP '19: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 130–138. DOI:https://doi.org/10.1145/3320435.3320449

[20] Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating critiquing-based music recommenders with conversational interaction. In *CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, NY, USA, 951–960. DOI:https://doi.org/10.1145/3357384.3357923

[21] Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. 2019. ContextPlay: Evaluating user control for context-aware music recommendation. In *UMAP '19: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 294–302. DOI:https://doi.org/10.1145/3320435.3320445

[22] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 13–21. DOI:https://doi.org/10.1145/3240323.3240358

[23] Michael Jugovac and Dietmar Jannach. 2017. Interacting with recommenders-overview and research directions. *ACM Transactions on Interactive Intelligent Systems 7*. DOI:https://doi.org/10.1145/3001837

[24] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *RecSys 2017 - Proceedings of the 11th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 229–237. DOI:https://doi.org/10.1145/3109859.3109873

[25] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*.

[26] Akiva Kleinerman, Ariel Rosenfeld, and Sarit Kraus. 2018. Providing explanations for recommendations in reciprocal environments. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 22–30. DOI:https://doi.org/10.1145/3240323.3240362

[27] B. P. Knijnenburg and M. C. Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook, Second Edition*. Springer US, 309–352. DOI:https://doi.org/10.1007/978-1-4899-7637-6_9

[28] Joseph A. Konstan and John Riedl. 2012. Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction 22*, 101–123. DOI:https://doi.org/10.1007/s11257-011-9112-x

[29] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *IUI '19: Proceedings of the 24th International*

Conference on Intelligent User Interfaces, Association for Computing Machinery, New York, NY, USA, 379–390. DOI:https://doi.org/10.1145/3301275.3302306

[30] 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 *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM Press, New York, New York, USA, 1–12. DOI:https://doi.org/10.1145/3290605.3300717

[31] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *Proceedings of The Web Conference 2020*, ACM, New York, NY, USA, 122–132. DOI:https://doi.org/10.1145/3366423.3380100

[32] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *WWW '20: Proceedings of The Web Conference 2020*, ACM, New York, NY, USA, 122–132. DOI:https://doi.org/10.1145/3366423.3380100

[33] Hongyu Lu, Ce Wang, Min Zhang, Feng Xia, Weizhi Ma, Yiqun Liu, Leyu Lin, and Shaoping Ma. 2019. Effects of user negative experience in mobile news streaming. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, New York, NY, USA, 705–714. DOI:https://doi.org/10.1145/3331184.3331247

[34] Malte Ludewig and Dietmar Jannach. 2019. User-centric evaluation of session-based recommendations for an automated radio station. In *RecSys 2019 - 13th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 516–520. DOI:https://doi.org/10.1145/3298689.3347046

[35] Valentina Maccatrozzo, Eveline van Everdingen, Lora Aroyo, and Guus Schreiber. 2017. Everybody, more or less, likes serendipity. In *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 29–34. DOI:https://doi.org/10.1145/3099023.3099064

[36] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 31–39. DOI:https://doi.org/10.1145/3240323.3240354

[37] Lei Mei, Liqiang Nie, Pengjie Ren, Jun Ma, Zhumin Chen, and Jian Yun Nie. 2018. An attentive interaction network for context-aware recommendations. In *CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, NY, USA, 157–166. DOI:https://doi.org/10.1145/3269206.3271813

[38] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *UMAP 2018 - Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 293–298. DOI:https://doi.org/10.1145/3213586.3225240

[39] Xi Niu, Fakhri Abbas, Mary lou Maher, and Kazjon Grace. 2018. Surprise me if you can: Serendipity in health information. In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery. DOI:https://doi.org/10.1145/3173574.3173597

[40] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, New York, USA, 1–13. DOI:https://doi.org/10.1145/3173574.3173677

[41] Puji Rahayu, D. I. Sensuse, B. Purwandari, I. Budi, F. Khalid, and N. Zulkarnaim. 2017. A systematic review of recommender system for e-portfolio domain. In *ACM International Conference Proceeding Series*, Association for Computing Machinery, New York, New York, USA, 21–26. DOI:https://doi.org/10.1145/3029387.3029420

[42] Darius A Rohani, Andrea Quemada Lopategui, Maria Faurholt-jepsen, Lars v Kessing, and Jakob E Bardram. 2020. MUBS : A Personalized Recommender System for Behavioral Activation in Mental Health. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. DOI:https://doi.org/10.1145/3313831.3376879

[43] James Schaffer, John O'Donovan, and Tobias Höllerer. 2018. Easy to please: Separating user experience from choice satisfaction. In *UMAP 2018*

- *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 177–185. DOI:https://doi.org/10.1145/3209219.3209222

[44] Tobias Schnabel, Paul N. Bennett, Susan T. Dumais, and Thorsten Joachims. 2018. Short-term satisfaction and long-term coverage: Understanding how users tolerate algorithmic exploration. In *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, Inc, New York, New York, USA, 513–521. DOI:https://doi.org/10.1145/3159652.3159700

[45] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Fairness in package-to-group recommendations. In *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 371–379. DOI:https://doi.org/10.1145/3038912.3052612

[46] Harald Steck. 2018. Calibrated recommendations. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 154–162. DOI:https://doi.org/10.1145/3240323.3240372

[47] Arthur A. Stone and John M. Neale. 1984. Effects of severe daily events on mood. *Journal of Personality and Social Psychology* 46, 1 (January 1984), 137–144. DOI:https://doi.org/10.1037/0022-3514.46.1.137

[48] Maria Stratigi, Jyrki Nummenmaa, Evaggelia Pitoura, and Kostas Stefanidis. 2020. Fair sequential group recommendations. In *SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing*, Association for Computing Machinery, New York, NY, USA, 1443–1452. DOI:https://doi.org/10.1145/3341105.3375766

[49] Taavi T. Taijala, Martijn C. Willemsen, and Joseph A. Konstan. 2018. MovieExplorer: Building an interactive exploration tool from ratings and latent taste spaces. In *SAC '18: Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, Association for Computing Machinery, New York, New York, USA, 1383–1392. DOI:https://doi.org/10.1145/3167132.3167281

[50] Chun Hua Tsai and Peter Brusilovsky. 2018. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *IUI '18: Proceedings of the 23rd International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, New York, USA, 239–250. DOI:https://doi.org/10.1145/3172944.3172959

[51] Chun-Hua Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, ACM, New York, NY, USA, 391–396. DOI:https://doi.org/10.1145/3301275.3302318

[52] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for recommender systems: The past, the present and the future. In *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 123–126. DOI:https://doi.org/10.1145/2959100.2959158

[53] Xuhai Xu, Ahmed Hassan Awadallah, Susan T Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding User Behavior For Document Recommendation. In *WWW '20: Proceedings of The Web Conference 2020*, ACM, New York, NY, USA, New York, NY, USA, 1–7. DOI:https://doi.org/10.1145/3366423.3380071

[54] Yuan Yao and F. Maxwell Harper. 2018. Judging similarity: A user-centric study of related item recommendations. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 288–296. DOI:https://doi.org/10.1145/3240323.3240351

[55] Michele Zanitti, Sokol Kosta, and Jannick Sørensen. 2018. A User-Centric Diversity by Design Recommender System for the Movie Application Domain. In *WWW '18: Companion Proceedings of the The Web Conference 2018*, Association for Computing Machinery (ACM), New York, New York, USA, 1381–1389. DOI:https://doi.org/10.1145/3184558.3191580

[56] Guoshuai Zhao, Hao Fu, Ruihua Song, Tetsuya Sakai, Zhongxia Chen, Xing Xie, and Xueming Qian. 2019. Personalized Reason Generation for Explainable Song Recommendation. *ACM Transactions on Intelligent Systems and Technology* 10, 4 (July 2019), 1–21. DOI:https://doi.org/10.1145/3337967

[57] Qian Zhao, F. Maxwell Harper, Gediminas Adomavicius, and Joseph A. Konstan. 2018. Explicit or implicit feedback? engagement or satisfaction?: A field experiment on machine-learning-based recommender systems. In *SAC '18: Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, Association for Computing

Machinery, New York, New York, USA, 1331–1340. DOI:https://doi.org/10.1145/3167132.3167275

[58] Qian Zhao, Martijn C. Willemsen, Gediminas Adomavicius, F. Maxwell Harper, and Joseph A. Konstan. 2018. Interpreting user inaction in recommender systems. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, NY, USA, 40–48. DOI:https://doi.org/10.1145/3240323.3240366

[59] Yong Zheng. 2017. Context suggestion: Empirical evaluations vs user studies. In *WI '17: Proceedings of the International Conference on Web Intelligence*, Association for Computing Machinery, Inc, New York, NY, USA, 753–760. DOI:https://doi.org/10.1145/3106426.3106466

Behaviours and Perceptions of Recommender Systems in Movie Streaming

Patrick Skov Johansen
Interaction Design
Dept. of Computer Science, AAU
Patrick.skov.johansen@gmail.com

ABSTRACT
From the recommender systems research literature, it is not evident whether current recommender systems provide good user experiences for its users. To investigate this, we conducted 14 semi-structured interviews about participants' usage and perception of the recommendations provided by recommender systems in movie streaming services. Our study provides evidence that people seek external information to justify a selection, recommendations that are perceived as opaque provide insufficient evidence to why it might fulfil the needs of the participants. Lastly, users rely on collaborative filtering recommendations when conversing with peers while finding content-based recommendations useful when browsing streaming services.

Author Keywords
Recommender systems; User experience; Explanations; Justification; Transparency

CSS Concepts
Information systems → *Recommender systems*
Human-centered computing → *Empirical studies in HCI*

INTRODUCTION
Recommender systems (RS) has been embedded in many modern technologies to personalise the user experience. LinkedIn regularly recommends professionals the user might know, and Facebook does the same with their friends. Central to the design of both services is the news feed, which they tailor to maximise engagement [13]. Recommendations also appear in leisurely services such as Netflix and Spotify, where the personalised experience drives satisfaction, retention, and engagement. RS have become ubiquitous [4], and hyper-relevance is the new norm [1].

RS research has long been focusing on algorithmic optimisation [3,26,27] on topics such as rating predictions of users [17], profit maximisation [5], and predicting online performance [21]. Recently, the literature has started to include investigations that focus on enhancing the user experience, for example by factoring the user's context into the recommendation [2,23,30] and providing explanations for the recommendations [15,19,22,29]. Kunkel et al. [19] studied the effects of personal and impersonal explanations and found the improvements in algorithmic accuracy next to meritless if RS fails to convey the rationale behind its decisions. These investigations related to the user experience are mostly novel prototypes and algorithms embodied in an unfamiliar interface; with specific tasks that users perform in short terms. The current body of knowledge lacks empirical investigations of recommender systems studied in the context of the users' lives. The extent to which current recommender systems in movie streaming deliver on the promises of alleviating choice overload, and improving the user experience, is not evident. The question remains, whether recommender systems are adopted and enrich the user experience for its users. With RS in a state of ubiquity, opportunities to conduct such investigations are apparent.

To address this question, we conducted semi-structured interviews with 14 users of movie streaming services. In our study, we found that the majority of participants actively seek external information to justify a selection. They were mostly avoiding the employed RS. Even when having to browse the libraries of the movie streaming services participants omit using the recommendations provided. Still, some participants with non-deviating tastes have good experience selecting recommendations when watching alone. All participants were insusceptible to opaque recommendations such as percentage recommendations and "Recommended for You" recommendations. As it turns out, participants need to be able to comprehend why an item might fulfil their psychological needs for them to consider it a viable option. Lastly, the participants were receptible both to content-based and collaborative filtering recommendation styles. Content-based recommendations appear in the streaming services, e.g. "Because you watched X", and were useful because they present something relatable. Users exchange recommendations using a collaborative filtering recommendation style when conversing with peers about movies and tv-shows. The recommendation exchange is based on the perception of their peers' taste, both when giving and receiving recommendations.

The main contribution of this article is three limitations of current recommender systems that affect their pragmatic quality and implications for how to overcome them:

1) *RS present insufficient information for users to evaluate the pragmatic quality,*
2) *RS fail to communicate the relevance of opaque recommendations, and*
3) *RS is unable to adapt to situational needs and contextual aspects.*

STUDY

We conducted 14 semi-structured interviews to investigate user's behaviours and perceptions of recommender system in movie streaming services such as Netflix. In preparation we developed an interview guide iteratively through a series of unstructured interviews about the topic. All 14 interview participants consented to let us record the interviews. The data was analysed thematically following the method by Braun & Clarke [7].

Developing the Interview Guide

In preparations for the semi-structured interviews, we conducted three unstructured interviews with fellow students to explore questions that people were able to answer based on recalled behaviour. The interviews were between 30-45 minutes. The unstructured interviews resulted in an interview guide containing four research topics with 3-4 associated questions. To answer the research topic questions, we translated them into interview questions that were suitable questions for the participants to answer. Each topic contained 4-5 interview questions. The interview guide was structured as follows:

1. Demographics – personal characteristics and movie streaming behaviour.
2. Selection behaviour – selection strategies, predictors for strategy selection, selection triggers.
3. Social recommendation behaviour – recommendation exchange, exchange characteristics, exchange considerations.
4. Perceptions and interactions with recommender systems – awareness, interaction, perceived control, scrutability.
5. Perception of recommendation quality – accuracy, diversity, novelty, usefulness.

An example of an interview question from the topic "Perceptions an interaction with recommender systems" is:

> *"What recommendations are you receiving at the moment?"*. With the follow-up question *"Has it always been those that are recommended to you?"*

Participants

Fourteen individuals (seven women and seven men) participated in the interviews. The participants were 19 - 52 years old (mean = 29,5). They use movie streaming services 3 - 7 days a week and had subscribed for 2 - 6 years (mean = 4,07). 12 of 14 was in a relationship, and 1 of 14 had small kids living at home. Every participant watched both movies and tv-shows and watched both alone and with others.

Participants were recruited through snowballing, where the first few interview participants were close relatives or friends, who recruited other participants outside our social circle. We presented the interview as an investigation of movie streaming behaviour to the participants.

#	Age	Streaming services	Typical weekly usage	Perceived library size (Netflix)
1	25	Netflix, Viaplay	Most days	250.000
2	24	Netflix, CMore	Every day	750.000
3	25	Netflix, Viaplay	Most days	1.000
4	25	Netflix	Most days	700
5	19	Netflix, Viaplay, HBO	Most days	400
6	34	Netflix, TV2 Play, Viaplay, HBO	Every day	4.000
7	52	Netflix, TV2 Play, PLEX	Every day	5.000
8	48	Netflix, Viaplay, HBO	Every day	3.000
9	38	Netflix, HBO	Few days	500
10	25	Netflix	Few days	1.500
11	25	Netflix, Viaplay, HBO	Few days	1.000
12	23	Netflix, Viaplay	Few days	300
13	27	Netflix, TV2 Play, Viaplay, HBO	Every day	20.000
14	23	Netflix, HBO, Amazon Prime	Every day	2.000

Table 1: The 14 participants were between 19 - 52 years old. All participants subscribed to Netflix and had been doing so for between 2 - 6 years. Participants used the services from 3 - 7 days a week. When asked to report their perception of the library size of Netflix, the answers were also quite varied.

Procedures

After recruiting the participants, we scheduled an appointment for the interview via text messages. The interview consisted of six phases. (1) Introduction & consent – before initiation of the interview, we greeted the interviewee, presented the subject of the interview, and asking for consent to record the interview. All participants gave us their consent. (2) Demographics – the interview commenced with demographic questions about the interviewee and their movie streaming consumption habits. (3) Selection Behaviour – We asked the interviewees to tell how they select what to watch under different circumstances. (4) Social Recommendations behaviour – We asked the interviewees to tell about interactions with other people regarding movie streaming, especially concerning recommendations. (5) Perceptions of RS – we asked the interviewees to reflect on their perception of RS in movie streaming, how they interact with it, and to what degree it guides their decisions. (6) Perceptions of Recommendation Quality – we asked interviewees to reflect on their satisfaction with the recommended content, and its accuracy. We concluded the interview with a summary of interviewees responses, allowing them to correct any misunderstandings. The interviews lasted 28 to 38 minutes.

We conducted the interviews via phonecall directed through a laptop computer. On the laptop, we used the Audio Hijack[1] application to record the audio of the interviewee and the interviewer on separate audio channels in the same MP3 file.

Analysis

The interview recordings were transcribed manually and imported into Dovetail[2] for analysing the data. We analysed the data thematical following the method presented by Braun & Clarke [7] consisting of the following steps: (1) Familiarising yourself with your data, (2) initial coding, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes (6) producing the report.

FINDINGS

In the following section, we report on the findings from conducting interviews with 14 participants about their movie streaming behaviour and how they perceive and interact with the recommender systems they provide. All participants had used streaming services regularly for more than two years, and everyone was a subscriber of Netflix. The participants used streaming services more during the weekends than during weekdays in terms of time spent watching, as well as more during winter than summer months. Since all participants subscribed to Netflix, we asked them to report their perception of Netflix's library size (i.e. how many individual titles are in the library, tv-show episodes excluded). Their answers varied significantly from 300 to 750.000 titles (Median = 1.750). Unofficial sources report that there were 2.480 movies as of Oct 2, 2018 [11] and 1.079 tv-shows as of Feb 17, 2020 [12] in the Danish Netflix library. A total of 2.559 titles, however, content is added and removed regularly. The following section is structured according to three themes we through the analysis: (1) External versus Internal Recommendation Reliance, (2) Perception and Understanding of Percentage Recommendations, and (3) Recommendation Styles: Content-based versus Collaborative Filtering.

External versus Internal Recommendation Reliance

Participants used two approaches to find a movie or tv show to watch: (1) pre-selection; relying on external sources to decide one an item *before* opening the streaming services, and (2) browsing; relying on internal recommendations provided by the streaming services to make a decision. The participants did not exclusively use a single approach, because the context for their usage varies (e.g. watching individually versus together) and effectiveness could be diminished (e.g. by using the same account for watching together and individually). Still, some participants favoured a particular approach if it was useful for them.

Pre-selection: Craving Reassurance

Ten participants stated that they sometimes rely on external sources (e.g. recommendation from peers and searching the internet for reviews and ratings) to find a movie or tv-show to watch before opening the streaming services. In that sense, they make a *pre-selection* before exposure to the internal recommender systems in the streaming services. When reasoning about this behaviour, participants described a desire for reassurance that the movie or tv-show would be a satisfactory choice.

> *P11: "Primarily, I choose to watch what others have recommended me. You know, recommendations, or a review I read somewhere in a newspaper or some other place."*

Pre-selection demanded varying effort before the participant opened the streaming applications with the intent of watching something. The amount of effort correlated with the participant's need for reassurance and some evidence fulfilled the need easier than others. An example was recommendations among peers. All participants acknowledged their involvement in the verbal exchange of movie or tv-show recommendations among friends or family. In the exchange, participants described a consideration for the recommendation receiver's taste when selecting who the receiver was and what to recommend. Similarly, when the participants received recommendations from peers, they considered the sender's taste and, in some instances, the quality of previous recommendations from that person. If the outcome of this evaluation was favourable (e.g. this person knows what I like and have given good recommendations previously), the participants were inclined to try the recommended item without further reassurance.

Participants also searched the internet, e.g. IMDb, for reviews, ratings, and top lists. This approach revealed popular or classic items belonging to a particular genre. The items were approved by other people and were only personalised to the extent that the participant knew what they

1 https://rogueamoeba.com/audiohijack/
2 https://dovetailapp.com

were looking after. Participants using this approach relied on stranger's recommendations for quality assurance. They often expressed a principal separation of the assumed good and bad items based on a particular rating (e.g. "below 7/10 I will not consider it").

P2: "I have tried those top-10 lists, top-10 comedies. However, there is rarely something interesting that I haven't watched already."

P10: "We use IMDb quite often to check how high the rating of a movie is. My boyfriend is very critical of movies. To satisfy him, it has to have at least seven on IMDb for him to consider it."

Another reason for pre-selecting an item to watch was publicity. Four participants expressed a desire to watch a particular item they had previously discarded because of increased publicity. These participants wanted to avoid feeling left out when conversing about the said item.

P9: "I don't know if it is an unconscious fear of feeling left out or something else, but it is like, you would want to be able to participate in the conversation. I don't want to lie and say that I didn't like it without having watched it. Then you are at square one if you haven't and don't know what it is about."

The majority of participants who pre-selected items to watch were in a relationship and watched primarily together with their partner. They did not want to spend excessive time browsing, since matching situational preferences was perceived to be cumbersome and time-consuming, often resulting in frustration. The recommendations in the streaming services were often perceived to be irrelevant either because: the participants used individual profiles; resulting in recommendations only fitting one party, or a shared profile; resulting in recommendations that are somewhere between both party's taste.

Browsing: More of the Same

Eleven participants stated that they sometimes relied on *browsing* through the streaming services to find a movie or tv-show to watch. Participants who primarily relied on browsing, use the internal recommendations and often referred to identifying items similar to what they had watched previously. Participants who used browsing secondarily refered to browsing by genre according to their mood.

P7: "When I get in the mood to watch something, I go to the living room to find something. (...) Most of the time I watch Sci-Fi, so it's kind of easy to pick a recommendation along those lines."

Participants who favoured browsing described a need to relax, by watching something, before considering what would be watched. Then, they browsed through the front page to find something appealing. Participants favouring this approach was characterised by having a non-deviating taste or liked a particular genre, that the recommender systems could detect and recommend. Participants often referred to the "Because you watched X"-shelves (rows of items) when describing what they did to find a movie or tv-show to watch. Similarly, one participant did not consider the recommendations on the front-page but went to a particular previously watched item to find similar items to that. The *browsing* approach was favoured when watching individually and having separate accounts if in a relationship.

When parties watched together, browsing was ineffective because the recommendations were not adjusted to the preferences of the other party who is now joining the session. The participants favouring the *browsing* approach express a degree of helplessness about considering the other party's preferences and finding an item that satisfies them both.

P7: "(...) but if I have to find something else. For example, when my wife wants to watch a comedy, I have a tough time because my preference system is not at all set up for comedies."

Participants who favour the *pre-selection* approach use *browsing* secondarily if pre-selection activates were skipped or in the mood for something else. Contrary, to those who favour *browsing*, these participants browsed by genre and avoided the recommendations on the front page. These participants described it as frustrating because the services seldomly provide the necessary reassurance by themselves.

P11: "I don't pick those that it recommends. Maybe I am more aware of what I'm going for. Of course, I have spent my time browsing and getting frustrated, but I rarely look at what it recommends to me."

Perception and Understanding of the Percentage Recommendations

Participants were aware recommendation provided in the streaming services. One of them is the percentage recommendations appearing on Netflix. Three of the participants were not aware of it, nine had noticed it but ignored it, and two considered it in their decision-making. Participants who noticed it were sceptical and confused about what it meant leading to abandonment:

P6: "I don't quite get the logic, I am aware that some of the items belong to the same theme, but it is not like I'm using it as a guiding principle 'Well I'm gonna trust it'. I am sceptical about why this item is 98% but the other movie that I've wanted to watch for a while is only 80%. Precisely that percentage is, I don't know… I notice it, but I don't think much of it. I certainly do not consider it in my decisions."

The percentage recommendations were presented on the detail page of every item in Netflix's library. The same page participants navigated to, to read descriptions and watch trailers. In the presented user interface, the percentage recommendation was prominently placed in the top left, highlighted in green, stating, e.g. "95 % Match". Whether the value was an output of algorithmic processing was not evident, but that was the assumption as of this writing. Participants ignored the recommendation, and instead rely on external sources such as ratings, reviews, and general publicity to justify their decisions:

> *P7: "(…) It is not something that I notice very much. Whether it is 91, 98 or 89% is not something that I… It is more other factors that make a difference. Was there, as I said, a rating of it. That would be preferable."*

The two participants who considered the percentage recommendation used it to 'tip the scales' whenever in doubt. They went to inspect the movie or tv-show and consulted the percentage if uncertain about their decision. Then, if the percentage is perceived to be high, it piques their interest. If not, then it is undoubtedly a sub-optimal decision:

> *P10: "I get totally fooled by it. I am completely susceptible, and I don't know why. If I have a little doubt if it fits my taste, then if it says 98 %, I have to try it out. But if it says 92% or something, then I don't feel the urgency to try it."*

> *P3: "I inspect the description and the percentage to make a decision. (…) If the percentage is below like 60%, I discard it."*

Even though the percentage recommendations were considered in decision-making, the participants did not adjust their perception of its accuracy after an experience. Participants had not reflected on how well the percentage reflected their satisfaction with the item. They simply stated that sometimes was accurate, while other times it did not make sense.

The percentage recommendation was presented as a type of explanation for why a particular item is presented. The "X % match" label does, however, not offer much insight into either the "what"; what does it match? Or the "how"; how does it match? The exception is the items presented in a shelf with the title "Because you watched X" misleading the user to perceive a content-based similarity comparison between the previously watched item and the new item. Whether this perception is appropriate was not apparent.

> *P10 "I have also noticed that thing that says '98 % similar to you' or 'something you have watched previously' or something along those lines. I have noticed that. (…) I think it is based on a comparison of the things I've already watched previously."*

Recommendation Styles: Content-Based versus Collaborative Filtering

Participants proactively searched for information that could justify why a particular item might be a satisfactory choice. The justification needed was varied among participants and their favoured approach. When participants relied on recommendations, two recommendation styles appeared to solve this need: (1) content-based recommendations and (2) collaborative filtering recommendations.

The *content-based* recommendation style appeared in the streaming services. It presented the participants with a previously watched item and recommended new content based on similarity to that item. Participants refeed to this as something relatable, relying on previous experiences to justify their decision. The *collaborative-filtering* was used when conversing with peers. It relied on similarities between peers to justify why a movie or tv-show should be a satisfactory choice. Participants created a perception of their peers' taste and used it to justify what to recommend, and whose recommendations should be tried. Participants rarely discussed the content of the movies of tv-shows explicitly, and the presentation was often an excited expression from the recommender.

Content-Based Recommendations

The participants who relied on the recommendations from the streaming services, frequently refered to the content-based recommendations presented with the title: "Because You Watched X".

> *P8:" If I watched a movie with superheroes, then it recommends something similar because I watched it. I think that is why I rely on it because it says, 'You watched this'. It presents me with something relatable."*

Another instance of content-based recommendations is after having watched a movie or the last episode of a tv-show. All participants had noticed these but rarely used them. They described that they seldomly continue watching item after item, making the content-based recommendations irrelevant in the situation they were presented. Two participants expressed curiosity about a recommended movie but did not immediately watch it. Instead, they looked it up on the internet later and considered it for another session.

When comparing the content-based recommendations to the other recommendations in the streaming services, such as the "Recommended for You"-shelves on the front-page, the content-based recommendations were more appealing. The "Recommended for You"-shelves did not receive much attention in the interviews, but it was not apparent whether the participants decide not to use them. The content-based "Because You Watched X" recommendations was a recurring theme in the interviews with participants who used the internal recommendations. As the quote above emphasises, it presented something relatable. The content-based similarity recommendations were the most useful of the internal recommendations.

Collaborative Filtering Recommendations

The content-based recommendation style stood in contrast to the ones employed in the verbal exchange of recommendations with peers. Here, a recommendation style more closely resembling *collaborative filtering* is used (a type of recommender algorithm built on the assumption that if people like the same items, they will continue to do so). When exchanging recommendations verbally with peers, the participants described a recommendation style that revolves around perceptions of their peers' tastes compared to their own. Here the content is an implicit consideration and seldomly presented.

> *P10: "I cannot help but consider their personality when judging if it is something they would like. And compare whether it would be something I would like as well. For example, I was recommended Outlander, I think it is called. In the beginning, I thought like 'Okay, I don't know if it would be something for me, but I will give it a try". Then I watched it and disliked it. But I gave it a try, even though my initial thought was bad."*

When participants conclude that they have a similar taste as the recommender, they often opt to try watching the item, and sometimes despite it being outside their taste.

> *P5: "I have a friend who recommended me Game of Thrones. At the time I hadn't watched it yet, but he said it was worth a try. So, I tried watching it and I was hooked. In other cases, I give the recommendation a try and figure out that I dislike it and stop watching. Based on their recommendation, I give it a try and stop watching if it turns out that I dislike it."*

The participants described three presentation styles when sharing recommendations: (1) "It is really good", (2) "You will like this", and (3) "This is something else!".

> *(1) P14: "We have a like a group chat, and someone just says, you know, 'I have watched three episodes. That's really good. You should watch it!'"*

> *(2) P11: "I get the most recommendation from people who know me well. They often say, 'I've seen this, and I just know you'll love it!'"*

> *(3) P12: "They remember them when people say like 'OMG this was so crazy'. For example, with Tiger King people were like 'This is just weird, you have got to see it', and then you can't deny the curiosity."*

All these presentation styles had a common absence of details about the content and similarity to previously watched items, in contrast to the participants' use of internal content-based recommendations. Another difference compared to the content-based recommendations is the development of status among recommenders. When sharing recommendations, participants expressed a reliance on particular individuals who had given satisfactory recommendations in the past.

> *P11: "I have to admit that it depends whether I think the person has given good recommendation in the past, where we have similar taste. I generally don't watch anything that I don't think is good."*

Contrary, participants who relied on recommendations from the streaming services were unable to describe how the different recommender systems compared.

DISCUSSION

Recommender systems have been a topic of research for more than 20 years. Traditionally, RS research focused on algorithmic optimisation but has more recently started to include investigations focusing on user experience. Current research efforts evaluate novel inventions with unfamiliar interfaces detached from its users' everyday lives. The current body of knowledge lacks empirical investigations of the uptake of recommender systems, its efficacy and users' experiences interacting with them. To address this problem, we conducted 14 interviews with experienced and frequent users of movie streaming services that provide recommendations. We found that the participants rely on external and internal sources to find a movie or tv-show to watch, they mostly disregard percentage recommendations that appear opaque, and are receptible to both content-based and collaborative filtering recommendation styles.

Limitations of Internal Recommendations

Our study showed that users of recommender systems actively avoid recommendations and find justification from external sources. We find this paradoxical since the many years of research improving accuracy and user experience has led to this situation where users omit using the tools, they are provided to make their choices easier. Even though browsing the vast libraries of content is frustrating, users disregard the recommendations supposed to be tailored to their preferences. Instead, users rely on external sources, seeking recommendations from strangers as a reassurance for a satisfactory choice. Our study shows that the efficacy of RS in movie streaming does not satisfy the needs of its users. When users browse for a movie or tv-show, they look for something specific as opposed to picking an item at random. There is an intent of watching as if a need has to be satisfied. The users look for information that will reveal the propensity for an item to satisfy that need. The item that is most likely, or likely enough, to satisfy the need will be selected. Now, when information, enabling the user to reason about an item's propensity to satisfy the need, is unavailable, our study shows that people disregard it and look for alternatives. We rely on Hassenzahl's [14] definition of user experience (UX), a *"momentary, primarily evaluative feeling (good-bad) while interacting with a product or service."*, to explain this dynamic in theoretical terms.

Hassenzahl [14] argues that the fulfilment of psychological needs (be-goals) is the driver of experience. A product's perceived ability to support such fulfilment denotes its hedonic quality. A product's perceived ability to support the achievement of "do-goals", such as watching a movie, denotes its pragmatic quality. Pragmatic quality is what facilitates the potential fulfilment of be-goals. Lacking pragmatic quality can impede the potential fulfilment of the be-goal, but the need itself may be unaffected. For example, a user may wish to be closer to their partner (be-goal) by watching a comedy movie together that they both will enjoy (do-goal). To watch a movie, they have to find one (do-goal). Despite having insufficient information to reason why a particular option has the propensity for the fulfilment of the be-goal, they still wish to fulfil it. Whether the user selects a recommendation is based on the pragmatic quality of the RS - the user's perception of the RS to support the achievement of the do-goal, i.e. its perceived ability to support the finding of a movie or tv-show that will satisfy them. If the pragmatic quality of the system is negative, i.e. getting frustrated with each other while browsing for a movie - it is not unreasonable that users avoid using the system and use alternative approaches consequently. To summarise, the pragmatic quality of the RS is what determines whether users will rely on it to make a selection or not. So, when there is insufficient information to make that determination, users will rely on other approaches.

Our study indicates that the factors influencing the pragmatic quality of RS in movie streaming were situational. In the following, we will outline the two inter-related characteristics that seem to affect the pragmatic quality evaluation of RS in movie streaming services: social watching context and preference continuity.

The social watching context is either together with others or individual. The pragmatic quality of RS is positive when the social watching context is individual, and the user wishes to have a similar experience to one they have had previously. In that situation, users rely on the content-based recommendations if the RS presents the right content-based comparison. When users watch with another party whose preferences or situational needs are different, the pragmatic quality of RS diminishes. Having a shared profile does not overcome this limitation. When groups make decisions together, each member plays different roles and exhibit different influence, as described by Cao et al. [8]. Groups do not make decisions as a single unit and make compromises unpredictably. As a result, recommendations appear random-like to either member.

We introduce the term "Preference continuity" to describe users whose preferences are unaffected by situational aspects. Our study showed that these users frequently make decisions based on content-based recommendations. These users want more of the same, and they describe their preferences in terms of a genre they prefer, such as Sci-Fi. For this type of user, the pragmatic quality of RS is positive because it consistently presents them with content that has fulfilled their be-goals associated with that type of content in the past. The content-based recommendations present a previously fulfilled be-goal that the user can recall and relate to, and on that basis, evaluate whether a similar experience will fulfil their current be-goal. Since their preferences are continuous, i.e. non-deviating, that is often the case. Contrary, users characterised by preference fluctuation are influenced by situational aspects. Our study shows that this type of user is affected by temporal aspects and their mood when making a selection. Factoring these aspects make the prediction of relevant content significantly more complicated because they are affected by the time of day and day of the week [10], daily events [28], and the weather [9].

Persuasive External Recommendations from Peers

Our study shows that some users regularly rely on external recommendations from peers when selecting movies or tv-shows and that these recommendations are both more memorable and more persuasive than internal recommendations. Not all participants in our study reflected this behavioural pattern, which aligns with recent literature. McInerney et al. [22] investigated how to personalise recommendation and explanations jointly. Their results indicate that user behaviour is dependent on the presented explanation of recommendations, which aligns with the finding in our study that some users rely on internal content-based recommendation and others are more receptible to external collaborative filtering recommendations. Similarly, Millecamp et al. [24] found interaction effects between personal characteristics, explanations, and interaction. These studies suggest that the presentation of recommendations significantly affects individual users' behaviour when interacting with recommendations. An interplay exists between user and recommendation, which varies between sessions. The study by Kouki et al. [18] supports this notion. They found that a user's personal characteristics affect both the preferred number of available explanations to justify a decision, as well as the most persuasive explanation style.

Contrary to our study, Kouki et al. [18] found that people rate the persuasiveness of socio-centric explanations (from peers) as less favourable compared to item-centric. Similarly, Berkovsky et al. [6] found "humanoid" presentations to be less favourable compared to genre-based and star-rating presentation styles. We propose three factors that might contribute to the disparity in findings: presentation, context, and the subject of study.

The presentation style used by Kouki et al. [18] and Berkovsky et al. [6] is not comparable to how the participants of our study described the verbal exchange of recommendations. The participants in our study described the exchange revolving around an implicit perception of each other's taste which they used to evaluate whether they should try the recommendation. The actual presentation of the recommendation is an expression of excitement or surprise, intended to spark curiosity in expectation of receiving feedback so the recommender could calibrate their perception. Kouki et al. [18] presented the recommendations as a preference statement, "Your friend Cindy likes U2". Berkovsky et al. [6] presented a static list of recommendations compiled by a random "humanoid" with whom the user had no relationship, which the participants of our study expressed was an intrinsic part of verbal recommendation exchange.

The context for the evaluation by Kouki et al. [18] and Berkovsky et al. [6] was a graphical user interface on a crowdsourcing platform. Our study included an investigation of exchange recommendations in real-life situations. Being isolated from a social interaction may change what recommendations users are receptible to. However, the study by Jin et al. [16] suggests that the social context of the user does not affect any aspects of the experience. Another contributor could be the ecological validity of the studies. An experimental setup on a crowdsourcing platform, may not induce comparable to real-life social needs, e.g. belongingness, and relatedness.

The subject of the studies also varies. Kouki et al. [18] study the effects of explanation styles on persuasiveness in movie recommendations. Berkovsky et al. [6] study trust factors in movie recommendations. In our study, we refrained from making pre-determinations of what factors influence user's behaviour concerning recommender systems and have instead, reported the subjective perceptions and recalled behaviour about interaction with recommender systems, which also poses limitations in itself.

The reason why participants in our study were persuaded to try novel content recommended by peers is not decisive. Still, some indications point to emotional need of belonging and being able to participate in conversations with other people. Leary & Cox [20] argue that the need for acceptance and belonging can explain a large proportion of human behaviour. They explain that this motive is so fundamental that the first premise for every theory of social or cultural behaviour could be that people "have a pervasive drive to form and maintain at least a minimum quantity of lasting, positive, and significant interpersonal relationships".

To summarize, our study provides evidence that the pragmatic quality of RS in movie streaming services varies significantly. Users characterised by preference continuity enjoy the recommendations from RS when watching individually, resulting in good UX. However, when watching with others, the pragmatic quality of RS is diminished. Recommendation for individuals become irrelevant and aggregated recommendations appear random-like. We have relied on Hassenzahl's [14] definition of user experience, to describe the interplay between pragmatic quality and hedonic quality in each of these situations. We argue that opaque recommendations provide insufficient information for users to evaluate the pragmatic quality of the recommended content. Contrary to our study, other researchers have found human presentation styles unfavourable. We dispute these findings and propose three factors that might cause the disparity. Findings from our work support Hassenzahl's [14] notion that "fulfilment of psychological needs is the driver of experience", and that if the user cannot link experience to fulfilment, i.e. if they cannot reason why something recommended would satisfy their be-goal, they will look for alternatives.

Implications for Design

In this article, we have described our study of perceptions and behaviours with recommender systems in movie streaming. We have presented three limitations of current RS that affects its pragmatic quality. Namely, *(1) RS present insufficient information for users to evaluate the pragmatic quality, (2) RS fail to communicate the relevance of opaque recommendations, and (3) RS is unable to adapt to situational needs and contextual aspects.* In this section, we will provide implications for how future recommender systems can overcome these limitations.

Bridge External and Internal Recommendations

Our study showed that movie consumption behaviour is related to external social activates and driven by the associated be-goals (e.g. being related, knowledgeable, belongingness). Therefore, we propose to bridge the gap between the external social context and the internal RS. Specifically, this could come about with personal recommendations of movies and tv-shows from peers inside the movie streaming services. Since users are persuaded by strong reactions and perceptions of each other's taste, collaborative filtering recommender systems should identify commonalities between peers and ask them for reactions of watched items. Then those reactions should be used to justify why a similar peer would want to watch that item. As more peers watch and react to the item, users become enticed to do the same by introducing belongingness motivation and directly presenting the means to fulfil it. Our study shows that it is of utmost importance that the user has a relationship to presented peers, to induce the right needs successfully.

Justify in Terms of Relatable Behaviour

Our study showed that users have to be able to reason why a particular option has the propensity to fulfil their situational need. Users neglect recommendations that violate this by providing insufficient justification. Justification in this context is not to be confused with transparency, reasoning about the inner workings of the algorithm. To overcome this limitation, we urge future designers of RS to provide explanations in terms users' preferences or previous behaviour, e.g. "You usually watch documentaries in the weekends" or "It has been a while since you watched X". Naturally, this justification has to be comprehensible and avoid causing information overload. Ambiguous percentages do not provide reassurance. They instead cause confusions as to what the percentage is based on, and whether that basis is compatible with the user's situational need. We realise that such proposal brings about challenges that are dependent on the chosen technology, but if recommendations are unable to be justified users will look for alternatives if available.

Adapt to Situational and Contextual Aspects

Data about the user's situational needs are not readily available to RS. Users turn to external sources because they can tailor their queries and pick what information will help them in making a selection. Such convergent functionality is not afforded in current movie streaming RS but could be helpful for users to make a selection. Eliciting situational aspects that affect the user's needs could either be approached implicitly; predicting the user's needs based on their interaction with the system while browsing, or explicitly; asking for the user's input directly. Specifically, we propose a content exploration design that relies on explicit feedback. The design proposal is a conversational interface presenting previously watched items that the user critiques based on their situational needs and context. The user provides feedback about a collection of item's propensities to be satisfactory. The recommender system adapts and provides new recommendations incrementally incorporating novel items. The interaction proceeds iteratively until the user finds a satisfactory option. Then, the question remains what the presentation should contain. Our study showed that users are reassured by different types of information ranging from ratings, and popularity to recommendations from strangers. One approach to overcome this is having users select the sources of information they usually rely on when making a selection. In that way, they participate in the personalisation of their recommender system. The study by Norton et al. [25] showed that people show increased appreciation for products they create, which might also be the case for personalised recommender systems.

CONCLUSION

In this article, we have reported on how users of movie streaming services perceive and interact with the recommender systems they employ. Our work contributes to the body of knowledge with three limitations of recommender systems that diminish their pragmatic quality depending on the context of the watching session. These contributions are a result of analysing subjective reports on the efficacy and user experience of recommender systems. The limitations span research topics related to presentation, justification, explanations, and contextual aspects. Lastly, we propose implications for future designers of recommender systems to consider when attempting to overcome the limitations in future systems.

REFERENCES

[1] Accenture. *Put Your Trust in Hyper-Relevance*.

[2] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems* 23, 1 (January 2005), 103–145. DOI:https://doi.org/10.1145/1055709.1055714

[3] Bushra Alhijawi. 2019. Improving collaborative filtering recommender system results using optimization technique. In *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, Association for Computing Machinery, New York, NY, USA, 183–187. DOI:https://doi.org/10.1145/3369114.3369126

[4] Xavier Amatriain and Justin Basilico. 2016. Past, present, and future of recommender systems: An industry perspective. In *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 211–214. DOI:https://doi.org/10.1145/2959100.2959144

[5] Amos Azaria, Avinatan Hassidim, Sarit Kraus, Adi Eshkol, Ofer Weintraub, and Irit Netanely. 2013. Movie recommender system for profit maximization. In *RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems*, ACM Press, New York, New York, USA, 121–128. DOI:https://doi.org/10.1145/2507157.2507162

[6] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to recommend? User trust factors in movie recommender systems. In *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, New York, USA, 287–300. DOI:https://doi.org/10.1145/3025171.3025209

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (January 2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[8] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. 2018. Attentive group recommendation. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Association for Computing Machinery, Inc, New York, NY, USA, 645–654. DOI:https://doi.org/10.1145/3209978.3209998

[9] Jaap J.A. Denissen, Ligaya Butalid, Lars Penke, and Marcel A.G. van Aken. 2008. The Effects of Weather on Daily Mood: A Multilevel Approach. *Emotion* 8, 5 (October 2008), 662–667. DOI:https://doi.org/10.1037/a0013497

[10] Boris Egloff, Anja Tausch, Carl Walter Kohlmann, and Heinz Walter Krohne. 1995. Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion* 19, 2 (June 1995), 99–110. DOI:https://doi.org/10.1007/BF02250565

[11] Finder.com. List of Netflix movies released in Denmark (updated daily). Retrieved June 3, 2020 from https://www.finder.com/dk/netflix-movies

[12] Finder.com. Complete List of Netflix Denmark TV Shows (updated daily). Retrieved June 3, 2020 from https://www.finder.com/dk/netflix-tv-shows

[13] Jaron Harambam, Mykola Makhortykh, Dimitrios Bountouridis, 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

RecSys 2019 - 13th ACM Conference on Recommender Systems, Association for Computing Machinery, Inc, New York, NY, USA, 69–77. DOI:https://doi.org/10.1145/3298689.3347014

[14] Marc Hassenzahl. 2008. *User Experience (UX): Towards an experiential perspective on product quality*. Retrieved June 2, 2020 from www.marc-hassenzahl.de

[15] J. L. Herlocker, J. A. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM Press, New York, New York, USA, 241–250. DOI:https://doi.org/10.1145/358916.358995

[16] Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. 2019. ContextPlay: Evaluating user control for context-aware music recommendation. In *UMAP '19: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, New York, NY, USA, 294–302. DOI:https://doi.org/10.1145/3320435.3320445

[17] Joseph A. Konstan and John Riedl. 2012. Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction 22*, 101–123. DOI:https://doi.org/10.1007/s11257-011-9112-x

[18] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, NY, USA, 379–390. DOI:https://doi.org/10.1145/3301275.3302306

[19] 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 *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM Press, New York, New York, USA, 1–12. DOI:https://doi.org/10.1145/3290605.3300717

[20] Mark R. Leary and Cody B. Cox. 2008. Belongingness motivation: A mainspring of social action. In *Handbook of motivation science.* The Guilford Press, 27–40. Retrieved June 6, 2020 from https://psycnet-apa-org.zorac.aub.aau.dk/record/2008-00543-002

[21] Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting online performance of news recommender systems through richer evaluation metrics. In *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 179–186. DOI:https://doi.org/10.1145/2792838.2800184

[22] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, New York, New York, USA, 31–39. DOI:https://doi.org/10.1145/3240323.3240354

[23] Lei Mei, Liqiang Nie, Pengjie Ren, Jun Ma, Zhumin Chen, and Jian Yun Nie. 2018. An attentive interaction network for context-aware recommendations. In *CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, NY, USA, 157–166. DOI:https://doi.org/10.1145/3269206.3271813

[24] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or not to Explain: the Effects of Personal Characteristics when Explaining Music Recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ACM, New York, NY, USA, 397–407. DOI:https://doi.org/10.1145/3301275.3302313

[25] Michael I. Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA effect: When labor leads to love. *Journal of Consumer Psychology* 22, 3 (July 2012), 453–460. DOI:https://doi.org/10.1016/j.jcps.2011.08.002

[26] Li Pu and Boi Faltings. 2013. Understanding and improving relational matrix factorization in recommender systems. In *RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems*, ACM Press, New York, New York, USA, 41–48. DOI:https://doi.org/10.1145/2507157.2507178

[27] X. Shi and Z. Lou. 2017. An improved similarity calculation algorithm used in news recommender system. In *ACM International Conference Proceeding Series*, Association for Computing Machinery, New York, NY, USA, 112–116. DOI:https://doi.org/10.1145/3173519.3173532

[28] Arthur A. Stone and John M. Neale. 1984. Effects of severe daily events on mood. *Journal of Personality and Social Psychology* 46, 1 (January

1984), 137–144. DOI:https://doi.org/10.1037/0022-3514.46.1.137

[29] Chun Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, New York, NY, USA, 391–396. DOI:https://doi.org/10.1145/3301275.3302318

[30] Yong Zheng. 2017. Context suggestion: Empirical evaluations vs user studies. In *WI '17: Proceedings of the International Conference on Web Intelligence*, Association for Computing Machinery, Inc, New York, NY, USA, 753–760. DOI:https://doi.org/10.1145/3106426.3106466