Proceedings of the

Joint Workshop on
Interfaces and Human Decision Making for Recommender Systems

September 16, 2016

In conjunction with the
10th ACM Conference on Recommender Systems
Boston, MA, USA

Edited by

Peter Brusilovsky, Alexander Felfernig, Pasquale Lops,
John O'Donovan, Giovanni Semeraro, Nava Tintarev,
Martijn C. Willemsen

Preface

As an interactive intelligent system, recommender systems are developed to give recommendations that match users' preferences. Since the emergence of recommender systems, a large majority of research focuses on objective accuracy criteria and less attention has been paid to how users interact with the system and the efficacy of interface designs from users' perspectives. The field has reached a point where it is ready to look beyond algorithms, into users' interactions, decision making processes, and overall experience. This workshop will focus on the aspect of integrating different theories of human decision making into the construction of recommender systems. It will focus particularly on the impact of interfaces on decision support and overall satisfaction.

The aim of the workshop is to bring together researchers and practitioners around the topics of designing and evaluating novel intelligent interfaces for recommender systems in order to: (1) share research and techniques, including new design technologies and evaluation methodologies (2) identify next key challenges in the area, and (3) identify emerging topics.

This workshop aims at establishing an interdisciplinary community with a focus on the interface design issues for recommender systems and promoting the collaboration opportunities between researchers and practitioners.

The workshop consists of a mix of nine presentations of papers in which results of ongoing research as reported in these proceedings are presented and one invited talk by Bart P. Knijnenburg on "*User-Tailored Privacy for Interactive Recommender Systems*". The workshop is closed by a final discussion session.

We thank all PC members, our keynote speakers, as well as authors of accepted papers for making IntRS 2016 possible. We hope you will enjoy the workshop!

Peter Brusilovsky, Alexander Felfernig, Pasquale Lops, John O'Donovan, Giovanni Semeraro, Nava Tintarev, Martijn C. Willemsen

September 2016

Organizing Committee

Workshop Co-Chairs

Peter Brusilovsky, School of Information Sciences, University of Pittsburgh, USA
Alexander Felfernig, Institute for Software Technology, Graz University of Technology, Austria
Pasquale Lops, Dept. of Computer Science, University of Bari “Aldo Moro”, Italy
John O’Donovan, Dept. of Computer Science, University of California, Santa Barbara
Giovanni Semeraro, Dept. of Computer Science, University of Bari “Aldo Moro”, Italy
Nava Tintarev, Bournemouth University, UK
Martijn C. Willemsen, Eindhoven University of Technology, The Netherlands

Program Committee

Robin Burke, DePaul University
Li Chen, Hong Kong Baptist University
Jaegul Choo, Korea University
Marco de Gemmis, University of Bari “Aldo Moro”
Michael Ekstrand, Dept. of Computer Science, Texas State University
Gerhard Friedrich, Alpen-Adria-Universitaet Klagenfurt
Franca Garzotto, Polimi
Mouzhi Ge, Universitaet der Bundeswehr Munich
Sergiu Gordea, AIT
Dietmar Jannach, TU Dortmund
Aigul Kaskina, University of Fribourg
Bart Knijnenburg, Clemson University
Fedelucio Narducci, University of Bari “Aldo Moro”
Denis Parra, Pontificia Universidad Catolica de Chile
Francesco Ricci, Free University of Bozen-Bolzano
Olga C. Santos, aDeNu Research Group (UNED)
Christin Seifert, Uni Passau
Luis Terán, University of Fribourg
Juha Tiihonen, University of Helsinki
Marko Tkalcic, Free University of Bolzano
Christoph Trattner, KMI, TU-Graz
Katrien Verbert, KU Leuven
Markus Zanker, Free University of Bolzano

Table of Contents

Invited presentation

Accepted papers

User-Tailored Privacy for Interactive Recommender Systems

Bart P. Knijnenburg
Clemson University
bartk@clemson.edu

Abstract

Privacy issues are an undying obstacle to the adoption of recommender systems, because recommender systems critically rely on their users to disclose information about themselves. While there exist several technical solutions to reduce the exposure of such personal information (e.g. client-side personalization, homomorphic encryption, *k*-anonymity), the concept of privacy is an inherently human attitude associated with the collection, distribution and use of disclosed data, and this disclosure itself is also a human behavior.

This talk discusses one particular human-centric solution to reduce users' privacy concerns in using recommender systems: User-Tailored Privacy. User-Tailored Privacy is an approach to privacy that measures users' privacy-related characteristics and behaviors, uses this as input to model their privacy preferences, and then provides them with adaptive privacy decision support. In effect, it takes the decision-supporting functionality of recommender systems, and applies it to users' privacy decisions.

The talk will revolve around the implementation and evaluation of User-Tailored Privacy in an interactive, demographics-based recommender system that gives healthy living advice. This system personalizes its recommendations based on the answer to a broad array of questions that range from innocuous (e.g. age, gender) to very sensitive (e.g. religion, sexual activity, household income, and savings). The system asks these questions in a sequential order, and recommendations are adapted to the user's answers on the fly. User-Tailored Privacy is implemented in the form of *adaptive request orders* that prioritize questions that are likely to benefit the recommender, but skips questions that the user is likely to deem too sensitive to answer. I will present the outcomes of a user experiment with 672 participants that tested several means of ordering the recommendations.

Investigating Mere-Presence Effects of Recommendations on the Consumer Choice Process

Sören Köcher, Dietmar Jannach, Michael Jugovac, and Hartmut H. Holzmüller
TU Dortmund University, Germany
firstname.lastname@tu-dortmund.de

ABSTRACT

In various application domains, recommender systems explicitly or implicitly act as *virtual advice givers*. They are not only used to filter large item sets or point users to unknown but relevant items, their recommendations can also help users to make a decision given a limited choice set. Such a system is usually considered effective if the users adopt the recommendations because, for example, the system's suggestions match their preferences or because they generally trust in the system's benevolence and competence.

With this work we aim to further explore the *persuasive potential* of automated recommendations. Our specific goal was to investigate whether the mere presence of a recommendation has effects on the user's choice process. We conducted two online studies in which participants received either no recommendation or a random recommendation for a given decision scenario. The obtained results showed that the pure existence of recommendations can, depending on the decision scenario, make users more confident in their choices and reduce choice difficulty. Furthermore, we observed that in both studies even random recommendations led to an anchoring effect as the participants' choices were measurably biased by the characteristics of the recommended item.

Keywords

Consumer Choice Behavior, Persuasiveness, Anchoring

1. INTRODUCTION

Recommender systems (RS) can serve different purposes for their users. They, for example, help users to locate relevant items within large item collections or support them in discovering additional items of interest outside their typical preference patterns. A somewhat less explored role of recommenders is their capability of serving as *virtual advice givers* in scenarios where users make decisions given a limited set of choices.

Such systems are often more interactive and can implement a number of persuasive cues to increase the users' confidence in their decision. Additionally, providers can employ the persuasive potential of such systems to "convince" users to choose a certain recommended option, e.g., by providing appropriate explanations or by helping them understand the relevant decision factors [7, 11, 19].

Previous works on this topic focus, for example, on analyzing the influence of specific decision-support functionalities, like explanations, on persuasiveness [8, 10]. In contrast, our work aims to examine if the *mere presence* of an arbitrary advice or recommendation has an effect on the user's decision making process. There are different reasons why we conjecture that such effects might exist: Recommender systems are omnipresent today and users might generally assume that such systems are benevolent and competent [12]. As a result, they might consider the recommendations in some form during their decision making process. If users are, in contrast, skeptical, the recommended items could at least serve as reference points when comparing the options. Finally, the recommended items could serve as *anchors* [17], which bias the users' decisions.

To investigate the existence of such effects, we conducted user studies in which the participants had to make purchase decisions on fictitious e-commerce shops. One participant group received one randomly chosen element from the choice set as a recommendation; the other group received no recommendation at all. We decided to rely on random recommendations in our studies as this allows us to rule out potential effects related to the (perceived) quality of the recommendations themselves. Besides the question if a randomly chosen recommendation can represent an anchor and bias the final user decisions, our expectation was that the *mere presence of recommendations has a positive effect on choice confidence*, e.g., because the users are given a reference point for their decision. Higher choice confidence might lead to higher choice satisfaction, which in turn is supposed to increase the users' intention to actually make a purchase [16].

In summary, our research questions are as follows:

- **RQ1:** To what extent has the mere presence of a recommendation an effect on the customer's decision process?
- **RQ2:** Can the characteristics of a recommendation serve as an anchor for decision making?

2. STUDY DESIGN

Research Model. Figure 1 shows our research model. The independent variables are the presence of a recommendation (RS) and the user's domain expertise (Dom. Exp.). We include the latter variable assuming that expertise may have an impact on the users' decision confidence (Dec. Conf.). We include choice difficulty (Ch. Difficulty) as a construct as we hypothesize that users – utilizing the recommendation as reference point – might focus on a subset of the items as choice set. In turn, lower choice difficulty should also lead to higher decision confidence. We measure choice difficulty with indicators variables that assess the degree to which making

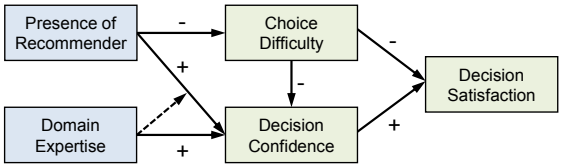


Figure 1: Research Model

the decision was perceived as (emotionally) challenging[1]. Finally, both decision confidence and choice difficulty are assumed to impact the users' decision satisfaction (Dec. Sat.).

Study Environment and Procedure. We created fictitious online shops for two different domains: backpacks (Study A) and hotels (Study B). In both studies, the scenario for the participants was that they were searching for an item to purchase, and we asked them to select one of the available items on the online site. The choice set sizes were 18 (backpacks) and 24 (hotels), respectively. In each case, additional item information was provided. We presented the weight, dimensions, volume, and price of the backpacks and the star category, community rating, distance to the city center and the price for the hotels. Half of the participants of each study received one *randomly* selected item as a recommendation, which was clearly marked as being a recommendation as sketched in Figure 2.

Figure 2: Fragment of the Fictitious Online Shop. Not all items are shown in the screen capture.

We recruited 164 and 239 participants for Study A and Study B, by distributing the URL of the online shop via email and on social network groups. The average age of the respondents was about 22 for both groups; more than two thirds were female participants. When accessing the website, the participants read the scenario and task description, were instructed to selected one of the options, and answered a post-task questionnaire. The participants were randomly assigned to the treatment groups.

3. RESULTS

The recommended item was actually chosen by 6.7 % of the participants of Study A, and by 3.8 % in Study B. These numbers roughly correspond to the theoretical chance that the recommended item was indeed the preferred option for a user. Simply displaying a random recommendation therefore *did not persuade* users to adopt the recommendations.

3.1 Structural Equation Modeling Results

We used Structural Equation Modeling (SEM) as an analysis instrument to detect relationships between the variables of our research model from Figure 1. Specifically, we used the PLS-SEM method, which is particularly recommended for this type of exploratory research for which no strong theory exists yet [9]. All constructs except the recommendation condition are measured with multiple questionnaire items.

3.1.1 Model Validity and Reliability

We applied different validity and reliability tests to our models and excluded indicators that were not reliably measuring a construct [9]. To check for *internal consistency* of the constructs, we measured *composite reliability* and *Cronbach's alpha* of our final model. The composite reliability values in both studies range from 0.88 to 0.95; Cronbach's alpha was between 0.79 and 0.93, i.e., all values were above the suggested minimum threshold of 0.7. To check for *convergent validity*, we calculated the AVE (Average Variance Extracted) value. The minimum AVE value across both studies was 0.68, which is again above the minimum threshold of 0.5. Finally, we verified *discriminant validity* by checking (a) that all cross loadings were smaller than the respective outer loadings and (b) that no squared variable correlations exceeded the AVE values of the respective constructs.

3.1.2 Observed Effects

In SEM models, *path coefficients* (β), which range from −1 and +1, express the strength of the relationships between two variables. The empirical t-values obtained through bootstrapping help us assess statistical significance. According to the literature [9], t-values above 1.96 indicate significance at the 5 % level, values above 2.57 at the 1 % level.

The middle columns of Table 1 show the β-values and t-values for the backpack study. The results confirm the hypothesized effects of the presence of a recommender on decision confidence and choice difficulty, which in turn both affect decision satisfaction. The main insight of this study is therefore that the mere existence of a random recommendation can (a) have a positive, statistically significant effect on the user's decision confidence and (b) lead to lower choice difficulty for users. This finding is relevant in practice as lower choice difficulty contributes to higher decision confidence ($\beta = -0.228$) and decision satisfaction ($\beta = -0.322$). Likewise, decision confidence is strongly tied with decision satisfaction ($\beta = 0.608$).

Table 1: SEM Results

	Backpacks		Hotels	
Path	β	t	β	t
RS → Dec. Conf.	0.15^{*}	1.97	$0.08^{\text{n.s.}}$	1.31
RS → Ch. Difficulty	-0.17^{*}	2.28	$-0.04^{\text{n.s.}}$	0.63
Dom. Exp. → Dec. Conf.	$0.02^{\text{n.s.}}$	0.22	$0.09^{\text{n.s.}}$	1.32
Dom. Exp. ×RS → Dec. Conf.	$0.10^{\text{n.s.}}$	0.84	$0.05^{\text{n.s.}}$	0.72
Ch. Difficulty → Dec. Conf.	-0.23^{**}	2.79	-0.37^{**}	6.72
Ch. Difficulty → Dec. Sat.	-0.32^{**}	5.97	$-0.06^{\text{n.s.}}$	1.23
Dec. Conf. → Dec. Sat.	0.61^{**}	11.59	0.67^{**}	16.55

Notes: β = Path coefficient with corresponding t-value, $^{**}p < .01$, $^{*}p < .05$, $^{\text{n.s.}}$= not significant.

[1]The questionnaire items can be found at http://ls13-www.cs.tu-dortmund.de/homepage/intrs13q.

The right-most columns of Table 1 show the results for the hotels. The obtained path coefficients indicate similar trends but the effects did *not* reach significant levels in this scenario. This suggests that the hypothesized effects depend on specifics of the domain or the decision scenario.

In both scenarios, the expertise of the users – in contrast to our expectations – had no measurable direct or moderating effect on decision confidence. This indicates that the observed mere-presence effects in the backpack domain applied equally to both experienced and less-experienced participants.

Overall, the results indicate that the mere presence of recommendations *can* measurably impact the user's decision process in terms of decision confidence and choice difficulty. However, while these effects were significant in the backpack domain, they did not reach significance in the hotel domain. Further research is therefore required to understand which factors cause these differences. One explanation could be that comparing item characteristics might be inherently easier in one of the domains. In fact, an analysis of the construct values related to choice difficulty revealed that choosing an item was considered significantly ($p < 0.05$) more difficult in the backpack domain, which could, e.g., be due to domain-specific trade-offs, such as the backpack's weight vs. its volume. This suggests that the presence of a recommender has more effects when the choice situation is more difficult. Another possible factor contributing to the perceived choice difficulty can be the size of the choice set [4]. In our studies, the choice set size was larger for the hotels than for the backpack domain. Nonetheless, the decision difficulty was perceived to be higher for the backpacks as mentioned above.

3.2 Analysis of Anchoring Effects

Besides the question to what extent a recommendation influences the decision making *process*, we aimed to investigate if the recommendations also had an effect on the *actual choices* of the participants. We have already mentioned above that participants did not blindly adopt the random recommendations. However, we hypothesize that the participants could (unconsciously) be biased in their final choice by the presented recommendation, i.e., that the recommendation served as an *anchor* for their decision. Specifically, we assume that participants select options that have *similar features* compared to the recommendation. To our knowledge, the existence of attribute-level anchoring effects has not been explored in the recommender systems literature so far.

Technically, we performed several univariate regression analyses to quantify to what extent the attributes of the chosen items were dependent on the features of the recommended item. Furthermore, as a simpler form of analysis, we calculated the correlations between the attribute values. The results of these analyses are shown in Table 2 and *clearly show the existence of anchoring effects for both scenarios.*

In the backpack domain, all features of the finally chosen item, i.e., weight, volume, and price, were positively and statistically significantly related to the recommended item. All correlation values (ρ) were also positive and significant at $p < 0.05$. Similar effects were observed for the hotel domain. On average, participants chose items that were similar to the recommended item in terms of the distance to the city center, the community rating, and the price. No anchoring effect was, however, observed for the star category in this domain. A possible explanation for this phenomenon could lie in the comparably coarse grained scale of the star category and that the participants might have already had a comparably strong mindset before the experiment regarding the star category of hotels they would possibly book.

Table 2: Result of the Anchoring Analyses

Backpack Scenario	β	t	ρ
Weight	0.257^{*}	2.573	0.278^{*}
Volume	0.146^{*}	2.127	0.233^{*}
Price	0.178^{*}	2.224	0.243^{*}
Hotel Scenario	β	t	ρ
Hotel Category	$0.028^{\text{n.s.}}$	0.381	$0.035^{\text{n.s.}}$
Distance from City Center	0.333^{**}	3.763	0.327^{**}
Recommendation Rate	0.299^{**}	3.762	0.327^{**}
Price per Night	0.202^{**}	3.106	0.275^{**}

Notes: β = Regression coefficient with corresponding t-value, ρ = Correlation coefficient, $^{**}p < .01$, $^{*}p < .05$, $^{\text{n.s.}}$= not significant.

To illustrate the strength of these effects, we looked at the item attributes and created different subsamples of the data (e.g., light vs. heavy backpacks). For example, when the system recommended a light backpack with a weight between 1.7 and 2.8 kilograms, the average weight of the chosen backpack was at 2.26 kilogram. When the weight of the "anchor" was higher and between 2.9 and 4.0 kilograms, the average weight of the selected backpacks went up by 13 % to 2.60 kilogram. Similar effect strengths were observed for other item features, which we find remarkable, given that the recommendations were randomly selected.

In an additional analysis we tested if domain expertise had an impact on the strength of the anchoring effect and incorporated these aspects into our regression models. We could, however, not observe any statistically significant main or interaction effects, which suggests that both novice and expert users seem to be equally susceptible to anchoring effects.

3.3 Research Limitations

Our research is mostly based on responses from students of our university. While the group is homogeneous and students are potential customers in both tested domains, we cannot state with certainty that the findings are representative for other societal groups. Furthermore, the participants did not actually make a purchase in the end, and our scenario was purely fictitious. On the other hand, as the participation was voluntary, no strong motivators exist for the participants to act dishonestly during the study.

4. PREVIOUS WORKS

Anchoring effects, as observed in both of our studies, were first discussed in the 1970s in the context of research on human judgment under uncertainty [17]. Anchoring means that people derive their final judgments or estimations for a given task using a heuristic that consists of adjusting a (possibly even arbitrary) initial value. Anchoring effects have been researched in different estimation and decision scenarios and, in particular in the Marketing literature, in purchase decision contexts, e.g., [13, 15], and [18].

In the RS literature only few works on anchoring effects exist. To what extent displaying predicted ratings for unfamiliar items influences the ratings assigned by users is discussed in [2] and [5]; how recommendations can impact

the users' willingness-to-pay is furthermore discussed in [1]. Anchoring effects on the item feature level, as reported in our work, have to our knowledge not yet been investigated.

In a broader context, anchoring effects can be seen as one of several possible approaches to implement *persuasive* recommender systems [19]. System-provided explanations are probably the most prominent approach in the RS literature to convince users to adopt a recommendation or make a certain choice, see, e.g., [8, 10]. Another approach to persuasion is to engage the user in the choice process, e.g., using an interactive product advisor, with the goal to promote certain items [20]. Finally, more deceptive means of persuasion include the manipulation of the recommendation list with the intent to exploit psychological phenomena like decoy effects [6].

In contrast to these works, our studies indicate that the mere presence of random recommendations can have a persuasive and biasing effect. More research is however required to understand the underlying reasons of these effects. Past research showed that users see (personalized) recommendations as a decision aid that can reduce the perceived effort and choice overload [3, 14]. The fact that even *random* recommendations are effective can have different reasons, for example, because users generally trust that such systems are benevolent and competent [12]. As an effect, users might feel *safer* with their choices when they are close to a recommended option.

5. SUMMARY AND CONCLUSIONS

Our work suggests that the mere presence of random recommendations can measurably affect the choice processes of users. In both studies reported in this paper we could observe anchoring effects on the attribute level, i.e., the participants exhibited a tendency to select items that had similar characteristics compared to the recommended reference item. In one of the tested domains, the presence of the recommender furthermore led to lower perceived choice difficulty and higher choice confidence.

Overall, our work therefore contributes additional evidence of the persuasive capabilities of recommender systems and their potential as decision-making aids. In terms of practical implications, the observed anchoring effects emphasize that recommenders can be valuable instruments for providers to guide the customer choice toward a desired direction. Since not all effects could be observed in both studies, more work is required to understand the underlying factors that determine the effective persuasiveness of recommendations in different scenarios.

6. REFERENCES

[1] G. Adomavicius, J. Bockstedt, S. P. Curley, and J. Zhang. Effects of online recommendations on consumers' willingness to pay. In *Proc. of Decisions@RecSys '12*, pages 40–45, 2012.

[2] G. Adomavicius, J. Bockstedt, S. P. Curley, and J. Zhang. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research*, 24(4):956–975, 2013.

[3] N. N. Bechwati and L. Xia. Consumers in cyberspace do computers sweat? the impact of perceived effort of online decision aids on consumers' satisfaction with the decision process. *Journal of Consumer Psychology*, 13(1):139 – 148, 2003.

[4] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. Understanding choice overload in recommender systems. In *Proc. of RecSys '10*, pages 63–70, 2010.

[5] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proc. of CHI '03*, pages 585–592, 2003.

[6] A. Felfernig, G. Friedrich, B. Gula, B. Hitz, T. Kruggel, G. Leitner, R. Melcher, D. Riepan, S. Strauss, E. Teppan, and O. Vitouch. Persuasive recommendation: serial position effects in knowledge-based recommender systems. In *Proc. of PERSUASIVE '07*, pages 283–294, 2007.

[7] A. Felfernig, G. Friedrich, D. Jannach, and M. Zanker. An integrated environment for the development of knowledge-based recommender applications. *Intl. Journal of Electronic Commerce*, 11(2):11–34, 2006.

[8] F. Gedikli, D. Jannach, and M. Ge. How should I explain? A comparison of different explanation types for recommender systems. *Intl. Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[9] J. Hair, G. Hult, C. Ringle, and M. A. Sarstedt. *Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage, 2013.

[10] J. Herlocker, J. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proc. of CSCW '00*, pages 241–250, 2000.

[11] B. Knijnenburg, M. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *UMUAI*, 22(4):441–504, 2012.

[12] S. X. Komiak and I. Benbasat. Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. *Information Technology and Management*, 5(1):181–207, 2004.

[13] K. C. Manning and D. E. Sprott. Multiple unit price promotions and their effects on quantity purchase intentions. *Journal of Retailing*, 83(4):411 – 421, 2007.

[14] V. Mindel. Choice anxiety in decision making: Why people turn to strangers for information. In *Proc. of ICIS '15*, 2015.

[15] J. C. Nunes and P. Boatwright. Incidental prices and their effect on willingness to pay. *Journal of Marketing Research*, 41(4):457–466, 2004.

[16] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proc. of RecSys '11*, pages 157–164, 2011.

[17] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[18] B. Wansink, R. J. Kent, and S. Hoch. An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research*, 35(1):71–81, 1998.

[19] K.-H. Yoo, U. Gretzel, and M. Zanker. *Persuasive Recommender Systems - Conceptual Background and Implications*. Springer, New York, 2013.

[20] M. Zanker, M. Bricman, S. Gordea, D. Jannach, and M. Jessenitschnig. Persuasive online-selling in quality and taste domains. In *Proc. of EC-Web '06*, pages 51–60, 2006.

Estimating party-user similarity in Voting Advice Applications using Hidden Markov Models

Marilena Agathokleous
Cyprus Univ. of Technology
30, Arch. Kyprianos str.
CY-3036, Limassol, Cyprus
mi.agathokleous@edu.cut.ac.cy

Nicolas Tsapatsoulis
Cyprus Univ. of Technology
30, Arch. Kyprianos str.
CY-3036, Limassol, Cyprus
nicolas.tsapatsoulis@cut.ac.cy

Constantinos Djouvas
Cyprus Univ. of Technology
30, Arch. Kyprianos str.
CY-3036, Limassol, Cyprus
costas.tziouvas@cut.ac.cy

ABSTRACT

Voting Advice Applications (VAAs) are Web tools that inform citizens about the political stances of parties (and/or candidates) that participate in upcoming elections. The traditional process that they follow is to call the users and the parties to state their position in a set of policy statements, usually grouped into meaningful categories (e.g., external policy, economy, society, etc). Having the aforementioned information, VAA can provide recommendation to users regarding the proximity/distance that a user has to each participating party. A social recommendation approach of VAAs (so-called SVAAs) calculates the closeness between each party's devoted users and the current user and ranks parties according the estimated 'party users' - user similarity. In our paper we stand on this approach and we assume that 'typical' voters of particular parties can be characterized by answer patterns (sequences of choices for all policy statements included in the VAA) and that the answer choice in each policy statement can be 'predicted' from previous answer choices. Thus, we resort to Hidden Markov Models (HMMs), which are proved to be effective machine learning tools for sequential and correlated data. Based on the principles of collaborative filtering we try to model 'party users' using HMMs and then exploit these models to recommend each VAA user the party whose model best fits their answer pattern. For our experiments we use three datasets based on the 2014 elections to the European Parliament[1].

CCS Concepts

•**Computing methodologies** → **Machine learning;**

Keywords

Hidden Markov Models; Voting Advice Applications; collaborative filtering; expectation maximization; recommender systems

[1]http://www.euvox2014.eu/

1. INTRODUCTION

Citizens, partly because of their lack of knowledge on the political issues, tend to avoid the democratic decision making process contributing in low voter turnout that affects the most advanced democracies. Ladner and Pianzola [18] specifically mentioned Switzerland, where the voter turnout does not exceed 50% by 1975. E-democracy tools and services can be used to inform people about the political stances of the parties (and/or candidates) who take part in the upcoming elections, aiming at increasing citizen participation and promoting direct involvement in political activities [22]. Voting Advice Applications (VAAs) are specifically designed e-democracy tools that further serve this purpose [17, 26]. They have been applied to facilitate citizens' decision making process by matching their political stances with those of parties and/or candidates. Findings have shown that VAAs' recommendations affect the decision making process of a significant part of voters, especially those who are undecided or belong to specific categories, such as people under 34 years old and/or first time voters [9, 26].

Recommender Systems (RSs) are software tools and techniques, which recommend products or services to users, in an effort to help them decide what they really need from the sheer volume of data that many modern online applications manage [14, 24]. Although the recommender systems are strongly affiliated with the field of e-marketing, several other application areas were also emerged. Recently, several researchers used recommender systems for e-elections in e-government to inform citizens about candidates and enhance their participation in democratic processes [7, 28], while Katakis *et al.* [15] introduced SVAAs (Social Voting Advice Applications), an extended form of VAAs that is based on the principles of collaborative filtering.

VAAs ask users and parties to fill a specific questionnaire that contains a number of policy statements, which are selected according to issues that concern the nation in time of elections and represent important political, economic and social issues [15, 19]. Figure 1 shows an example of such a policy statement along with the set of possible answers a user can select. The recommendation process that a VAA traditionally follows contains two main steps: first, it calculates the similarity scores utilizing the user's and the parties' and/or candidates' answers in the policy statements and then, the VAA ranks the parties according to party-user 'similarity'. Figure 2 presents an example taken from the German VAA of the elections to the European parliament in 2014.

Researchers from different research fields deal with many

Figure 1: A question that was included in EUVox 2014 along with the given set of answer options.

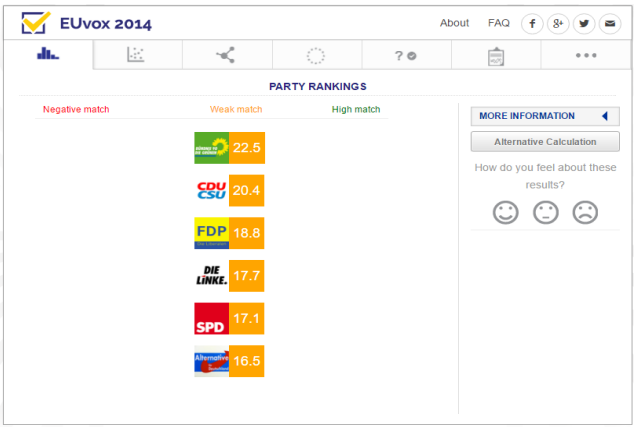


Figure 2: Party ranking based on party-user similarity as computed in traditional VAAs (EUVox 2014, Germany data).

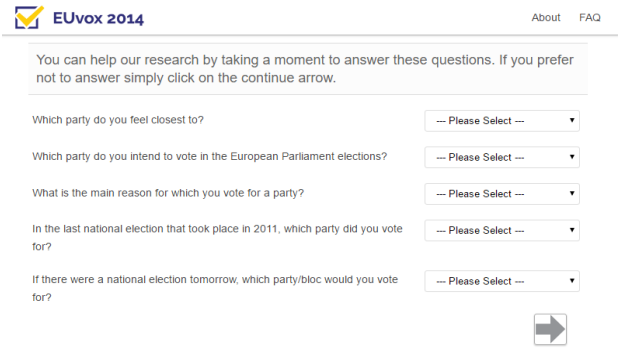


Figure 3: The supplementary questions as they appear in EUVox 2014.

aspects of VAAs [25]. Some of them investigate whether VAAs urge citizens to vote and whether recommendations made by these systems affect the final vote decision [9, 26]. Other researchers are interested in the design of VAAs dealing with practical issues such as the derivation of optimal party-user similarity estimation methods that accurately predict users' voting intention [20, 21, 29]. We note here that the estimation of similarity between users based on their choices from a set of products is a core problem in Recommender Systems as well.

Recently Katakis *et al.* [15] coined the term 'Social VAAs' (SVAAs) in an effort to describe VAAs, whose recommendation is based on the collaborative filtering philosophy that is widely used in RSs [12, 13]. SVAAs in addition to parties' answers to the policy statements, they also utilize models that capture the behavior - in respect to the policy statements - of each party voters. Thus, a social VAA has the same policy questionnaire with the traditional VAA but also party voters models created by estimating the joint probability of answer patterns and vote intention of each user. Vote intention is an opt in question which is included in VAAs as one of the supplementary questions. An example of supplementary questions included in VAAs is shown in Figure 3, where the vote intention question is the second one.

In SVAAs users are classified into groups according to their voting intention, i.e., party or candidate choice, and then models are created for each party to 'show' the common way, if any, in which party supporters fill the online questionnaire producing their own answer pattern. Then, the SVAA recommends new user with the party or the candidate whose users' model matches better their answer patterns. Figure 4 presents an example of the matching scores presented to a user based on the SVAA philosophy. SVAAs proved to make better voting predictions than the traditional matching schemes between users' and parties' profiles [1]. In addition, as recorded by users' feedback through the emoticons shown in the right part of Figures 2 and 4, SVAA recommendation surpasses VAA recommendation in terms of users satisfaction [6].

In order to tackle the recommendation problem of SVAAs, machine learning techniques [2] can be used to indicate the likelihood that a user belongs into a class, where each class corresponds to a specific party. In essence, what is accomplished with machine learning is to model each party cording to its supporters' answer patterns to policy statements. Thus, if a user is classified into a party, it is more likely this user has the same political positions with people who are already classified to the same party. Katakis *et al.* [15] resorted to clustering and classification approaches for generating vote advice in SVAAs and they showed that party voter modeling based on data mining classifiers and Support Vector Machines, achieve the best performance.

Tsapatsoulis and Mendez [30] dealt with building party voter models for SVAAs based on the probability to vote each one of parties participating in the German elections in 2013. They compared a Mahalanobis Classifier, a Weighted Mahalanobis Classifier and function approximation approaches, and they concluded that there is no much gain when using the probability to vote instead of the vote intention. They also noticed that non-linear party modeling techniques, such as neural network based ones, outperform the linear methods like Mahalanobis.

Tsapatsoulis *et al.* [29] in an effort to provide practical design guidelines for SVAAs dealt with the problem of finding the minimum number of VAA users required to build effective party's voter models. They limited their analysis

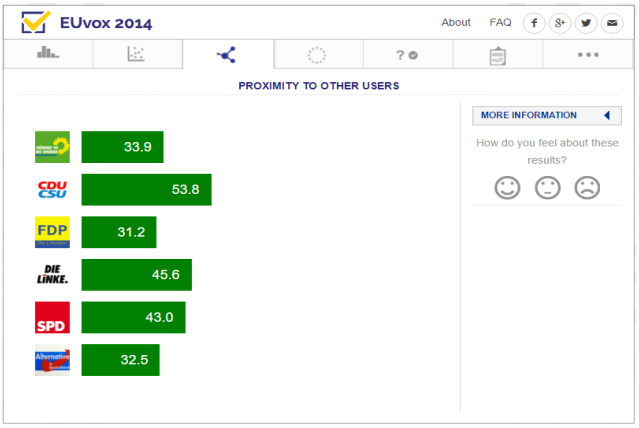


Figure 4: Party ranking based on matching scores between party models and user's answer pattern.

to the Mahalanobis Classifier for minimize the factors influencing their research questions. They found that, as the number of parties modeled is increased the performance of recommendation is decreased. In addition they showed that effective party voter models can be built based on a rather small number of user profiles.

In this work we adopt the social approach of VAAs and we investigate the application of Hidden Markov Model (HMM) classifiers for party-user similarity estimation in an effort to improve the effectiveness of social vote recommendation. HMM classifiers provide a way to apply machine learning to data represented as a sequence of correlated observations [2].

In VAAs the order in which policy statements are displayed to users is not important; however, policy statements are usually correlated and grouped into categories (e.g., external policy, economy, society, etc). Thus, opting from the various answer choices in each policy statement is related with selections in previous and subsequent policy statements. Given that the order of policy statements is kept fixed within each VAA one can assume that (a) answer patterns, that are sequences of choices for all policy statements included in the VAA that characterize 'typical' voters of particular parties can be found, and (b) the answer choice in each policy statement can be 'predicted' from previous answer choices. When users answer the questions, they are incrementally producing a sequence of symbols. Whenever a process includes a sequence of dependent observations, HMM classifiers can be used to model input sequences as generated by a parametric random process. This is our basic rationale for employing HMMs for obtaining similarity matching between parties and users for SVAAs.

We assume that VAA users, who support the same party, produce similar sequences of symbols, i.e., answer patterns. Thus, HMM classifiers can be used to predict and identify the 'path' that users, who support the same party, follow to answer the online questionnaire, and to create simple and compact models for each party, so as to be able to classify new users into the most probable party class. Although there is enough evidence about the appropriateness of HMM classifiers for SVAA recommendation, they have not been applied so far. This is probably due to the fact that there are simpler machine learning techniques that can be used instead. However, we strongly believe that HMMs have an advantage compared to other machine learning methods: they can capture the correlation between answers in different policy statements.

In short, the purpose of our paper is to introduce an SVAA method for similarity matching between parties and users based on HMMs and investigate its performance based on the accuracy of predicting their voting intention. We show that, even if the order in which the questions are answered in a VAA does not really matter, the HMM classifier performs quite well in estimating vote intention of unseen users. Nevertheless, the HMMs' performance relies on the smooth distribution of samples per party and on the consistency between the answers of the users, who are classified as belonging to these parties. Therefore in the cases where these conditions are not met, the results may not be satisfactory; in such case datasets used for training should be cleaned using outlier and/or rogue detection techniques [5].

To the best of our knowledge this is the first time HMMs are used to compute party-user similarity either in VAAs or in SVAAs. For our experiments we use three datasets derived from EUVox 2014. EUVox is an online application that was sponsored by the Open Society Initiative for Europe (European Elections 2014) and the Directorate-General for Communication of the European Parliament (area of internet-based activities/online media âĂŞ 2014). It purpose was to help voters to have quick access to information related to the political positions of the parties participated in the 2014 elections to the European Parliament (see more information at http://www.euvox2014.eu/). The chosen datasets differ in size, in the number of parties participating in the elections and in the population's distribution percentage among the various parties. An important, possible, contribution to researchers belonging to the Recommender Systems community is that the corresponding datasets, as well as many other VAA datasets, are freely available through the Preference Matcher Website[2]. One of the aims of the current work is to mobilize researchers of RSs to investigate the performance of their techniques on VAA data.

2. PROBLEM FORMULATION

The basic aim of a traditional VAA is to recommend parties to users. In such a case there is a set of N users $X = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_N}\}$, a set of U policy statements $Q = \{q_1, q_2, \ldots, q_U\}$, and a set of D political parties or candidates $P = \{\vec{p_1}, \vec{p_2}, \ldots, \vec{p_D}\}$. Each user $\vec{x_j} \in X$ and each political party $\vec{p_i} \in P$, has answered each policy question $q_k \in Q$.

Based on their answers, every political party or user can be represented in a vector space model:

$$\vec{x_j} = \{x_{(j,1)}, x_{(j.2)}, \ldots, x_{(j,k)}, \ldots, x_{(j,U)}\} \quad (1)$$

$$\vec{p_i} = \{p_{(i,1)}, p_{(i,2)}, \ldots, p_{(i,k)}, \ldots, p_{(i,U)}\} \quad (2)$$

where $x_{(j,k)}, p_{(i,k)} \in L$ are the answers of the j-th user and i-th party, respectively, to the k-th question. The vectors $\vec{x_j}$ and $\vec{p_i}$ are, usually, named user and party *profiles* respectively.

A typical set of answers is a 6-point Likert scale: L ={1 (Completely disagree), 2 (Disagree), 3 (Neither agree nor

[2]http://www.preferencematcher.org/?page_id=18

disagree), 4 (Agree), 5 (Completely agree), 6 (No opinion)}. In several cases, and in the majority of SVAA methods proposed so far, the sixth point is not taken into consideration since it does not correspond to a particular stance and is usually replaced with the third point, i.e., with 'neither agree nor disagree'. In this work we decided to keep the sixth point as a distinct emission symbol (see also Section 3) in order to avoid a common criticism by political scientists who strongly argue about the difference between these two categories [3]. As a result the set L, in the context of this study, becomes: L ={1,2,3,4,5,6}. Figure 1 shows an example of the way the policy statements in the EUVox 2014 appear and how the answer options are presented to VAA users.

The VAA recommendation task tries to approximate the unknown relevance $h(j,i)$ of user j to party i given the user's answers $\vec{x_j}$ and then to suggest a ranking of political parties based on user-party similarity. In machine learning terms, the task is to approximate the hidden function $h(j,i)$ with a function $\hat{h} : \mathbb{R}^U \times \mathbb{R}^U \to \mathbb{R}$, where $\hat{h}(\vec{x}_j, \vec{p}_i)$ is the estimation of the relevance of user j with political party i. Typically $\hat{h}(\vec{x}, \vec{p}) \in [0,1]$. In each case, the top suggestion p_q^j for user j should be:

$$p_q^j = \underbrace{argmax}_{i}(\hat{h}(\vec{x_j}, \vec{p_i})) \tag{3}$$

In many VAAs, the users are asked to answer a number of supplementary questions in addition to the U policy statements. One of these supplementary (opt in) questions is the vote intention of user i.e., which party the user intends to vote in the upcoming election. An example of the type of supplementary questions and how they appear in the EUVox 2014 is shown in Figure 3.

The main idea behind the SVAA is to use the vote intention variable y_j and model each party's voters using statistical or machine learning approaches. Thus, for every party i a model $\vec{M}_i$ is created using as training examples the subset $\mathbf{T_i}$ of user profiles who expressed voting intention for party i, that is $\mathbf{T_i} = [\vec{x}_j | y_j{=}i]$. Then, these models can be exploited to provide a recommendation based on collaborative filtering [11] that takes advantage of a VAA's voter community. In this case the top recommendation p_q^j for user j is given by:

$$p_q^j = \underbrace{argmax}_{i}(\hat{h}(\vec{x_j}, \vec{M}_i)) \tag{4}$$

In this work we use Hidden Markov Models to create the party-voter models $\vec{M}_i$ (see Section 3). Thus, Eq. 4 becomes:

$$p_q^j = \underbrace{argmax}_{i}(\hat{h}(V^j, \lambda_i)) \tag{5}$$

where V^j is the set of observations corresponding to user profile $\vec{x_j}$ and λ_i is the party-voters model for party i created using HMM training. The solution of Eq. 5 is obtained with the aid of Viterbi algorithm as usually happens in HMM classifiers [2].

An HMM is a double stochastic process that models data evolving in time. It is defined by a latent Markov chain, which consists of a finite number of states, and a number of observation probability distributions for each state. At each discrete time instant, the system switches from one state to another, while an observation is produced by the probability distribution according to the current state [16]. In an HMM, the states are not observable, i.e., they are 'hidden', but an observation is generated as a probabilistic function of the state, when the system visits the state [2].

An HMM is described by three parameters: $\lambda = (A, B, \pi)$, which can be estimated based on specialized Expectation Maximization (EM) techniques, such as the Viterbi or the Baum-Welch algorithm. The parameters are calculated through several training iterations, by using the entire training data set at each time, until an objective function is maximized. To avoid knowledge corruption, the data should be storage in memory and be trained from the start at each iteration, a costly and time consuming process. Therefore in real life, the datasets used for training HMMs are often small and this might significantly reduce their performance since the effectiveness of HMMs depend heavily on the availability of a sufficient quantity of representative training data to calculate the model parameters [16].

As already stated, in this work we try to optimize SVAA recommendation with the aid of a Hidden Markov Model classifier. This is, probably, the first time the HMMs are used in SVAAs and one of the very few times used in Recommender System applications in general. A possible explanation is the fact that within a VAA, and in many RSs, the observations corresponding to user (answer) choices are not time dependent. However, as we already mentioned, in VAAs user answer choices can be considered as a sequence of *correlated* observations while HMM states could correspond to the set of permissible answer options ('Completely disagree', 'Disagree', 'Neither agree nor disagree', 'Agree', 'Completely agree'). Under these circumstances the HMMs can be applied to VAA, as we have a sufficient number of states and a fairly rich set of data.

3. METHODOLOGY

An HMM is characterized by [2]:

- A set of W discrete states $S = S_1, S_2, S_3, ..., S_W$, with $G = g_1 g_2 ... g_T$ to be the state sequence (i.e., if we have $g_t = S_i$ that means at time t the system is in state S_i).
- A set of E observations $V = v_1, v_2, v_3, ..., v_E$, with $O = O_1 O_2 ... O_T$ to be the sequence of observations corresponding to states G.
- A state transition matrix A, that shows the probability of going from state S_i to state S_j: $A \equiv [a_{ij}]$ where $a_{ij} \equiv P(g_{t+1} = S_j | g_t = S_i)$.
- An observation emission matrix B, that describes the probability of observing v_e in state S_j: $B \equiv [b_j(e)]$ where $b_j(e) \equiv P(O_t = v_e | g_t = S_j)$.
- The probability distribution of being in the first state of a sequence: $\pi \equiv [\pi_i]$ where $\pi_i \equiv P(g_1 = S_i)$.

In our implementation we consider HMMs with three states, i.e., $W = 3$, $S = \{S_1, S_2, S_3\}$, labeled as S_1: 'Negative', S_2: 'Neutral', and S_3: 'Positive' corresponding to answer choices S_1: (Completely disagree, Disagree), S_2: (Neither agree nor disagree, I have no opinion), and S_3: (Agree, Completely agree) that could be given in the U policy statements of the

VAA questionnaire. Furthermore, there are six possible observations $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, where v_1: 'Completely disagree', v_2: 'Disagree' v_3: 'Neither agree nor disagree', v_4: 'I have no opinion', v_5: 'Agree', and v_6: 'Completely agree'.

Every state sequence G has length equal to the number of policy statements, i.e., $T = U = 30$ while the mapping from a user profile $\vec{x_j}$ (see also Eq. 1) to an emission sequence $V^j = \{v_1^j, v_2^j, v_3^j, ..., v_E^j\}$ is obtained as follows:

$$v_q^j = x_{(j,q)} + |L| \cdot (q-1) \quad (6)$$

where $x_{(j,q)}$ is the answer choice of user j to policy statement q ($q = 1, ..., E$), L is the set of answer options (see also Section 2) and $|L|$ is its cardinality, i.e., the number of answer options in the policy statements. Thus, in our case $|L| = 6$.

As an example consider that a VAA user selected 'Completely Disagree' in the 1st policy statement; then, according to Eq. 6 the recorded observation in the 1st place of the sequential answers of the voter would be: $1 + 6 * (1 - 1) = 1$; whereas if the answer choice in the 23rd policy statement was 'I agree', then the observation $4 + 6 * (23 - 1) = 136$ would be registered in the 23rd place of the V^j sequence.

An HMM is fully described by three parameters: $\lambda = (A, B, \pi)$. In the framework of this work we consider that each party voters can be modeled by an HMM λ_i since the way VAA users respond to the first policy statement differs among supporters of different parties reflecting into different π_i, the same holds for any other policy statement reflecting in different B_i, while the way answer choices are given in two consecutive policy statements also varies among different party supporters reflecting into different A_i.

4. EXPERIMENTAL RESULTS

4.1 Datasets

As in the majority of VAA and SVAA methods, in this work we set the performance criterion to be the accuracy of predicting a user's vote intention. This also aligns with the approach followed in Recommender Systems where the criterion is the accuracy of predicting users' ratings. Thus, we carried out experiments to measure the performance of voting prediction by applying the HMM classifier on three EUVox datasets derived from Denmark, Bulgaria and Czech Republic. EUVox is an EU-wide voting advice application that was utilized during the 2014 European Parliament elections. Its questionnaire consists of 30 policy statements and it is based on European-wide issues, issues that are salient for voters in a particular region, and country-specific issues. The policy statements are clustered into three groups; to those that refer to European Union issues, to those dealing with economy, and to those related to societal issues.

The three datasets were chosen such as to differ in size. The number of samples of the Bulgarian dataset is quite small; approximately 2800 entries were correct and also contained a voting intention answer. The Czech dataset is approximately 5 times larger than the Bulgarian while the Danish dataset is the largest; it contains almost 4 times more samples than the Czech dataset. In addition the number of parties participating in the elections varies among the selected datasets while the same holds for the population distribution among the various parties. The Danish dataset is characterized by a rather smooth distribution of samples per party which is not the case in the Bulgarian and Czech datasets (see Figure 5). These differences helped us to examine the behavior of HMMs when there is no sufficient number of data points per party and when the number of samples varies among parties.

In order to measure the performance of voting prediction using HMMs, we took into consideration only the users who expressed a voting intention for a specific party. Therefore, the questionnaires of the users, who did not answer the supplementary question on voting intention, or answered either 'not decided yet' or 'I will not vote' were exempted. In all three datasets approximately 40% of the VAA users expressed voting intention for a specific party. The main characteristics of the used datasets are summarized in Table 1.

4.2 Results and Discussion

Experiments were designed to investigate the performance of social voting recommendation using HMMs for estimating party-user similarity. For the evaluation we divided the users of dataset into a training and a test set [8]. A HMM is built against the training set $T_r = \{(\vec{x}_j, y_j) | j = 1...N_l, y_j \neq \emptyset\}$ consisting of the profile vectors $\vec{x}_j$ corresponding to user answers to the online questionnaire along with the user's expressed vote intention y_j. Evaluation of the trained HMMs on unseen data was facilitated using the test set $T_e = \{(\vec{x}_t, y_t) | (\vec{x}_t, y_t) \notin T_r, t = 1...N_t, y_t \neq \emptyset\}$ which is a set of profiles and voting intention pairs ($\vec{x}_t$, y_t) not used in the training set.

In order to perform our experiments we resorted to Matlab's HMM toolbox. This toolbox was built by Kevin Murphy and it uses the Baum-Welch (BW) algorithm for estimating the parameters of HMMs with discrete outputs [23]. We created an HMM $\lambda_i = (A_i, B_i, \pi_i)$, for every party included in each one of the datasets. Thus, we ended up with seven HMMs for the Danish and Bulgarian datasets and ten models for the Czech dataset. After training the party models using the training set T_r the test set T_e was used to classify unseen users, expressed through their profiles, into the party in which the user most likely belongs to, i.e., the user's answer pattern most accurately fits i-th party's model. In the end, to examine the voting prediction performance of HMMs, the real voting intention of each user in the testing set was compared to the predicted voting intention, that is the party id of the party in which they were classified. At the end an overall score of how well the algorithm performed was calculated using the Precision, Recall and F-measure scores and then a total weighted average was estimated [29].

In Tables 2-4 we can see the results for each party of the three datasets while Table 5 shows the total weighted averages for Precision, Recall and F-measures, and the Mean Average Precision (MAP) for each dataset. The aggregate results of HMMs obtained in the Danish and Czech datasets are better than the ones obtained in the Bulgarian dataset, but without a marked difference. The HMM classifier achieved a similar overall prediction performance for the Danish and Czech datasets, with the former to be slightly better.

In the Danish dataset the smooth distribution of samples per party (see Figure 5(a)) along with the homogeneity of answer patterns among the supporters of the same party reflects in quite smooth performance across parties as it can be seen in Table 2. However, the prediction performance for the sixth party, which holds the majority of the users, exceeds the performance of the others. The third and the

Table 1: Datasets' characteristics

Dataset	# samples (Questionnaires)	# samples in the training set	# samples in the test set	# parties modeled
Danish	53284	31970	21314	7
Bulgarian	2755	1653	1102	7
Czech	15278	9167	6111	10

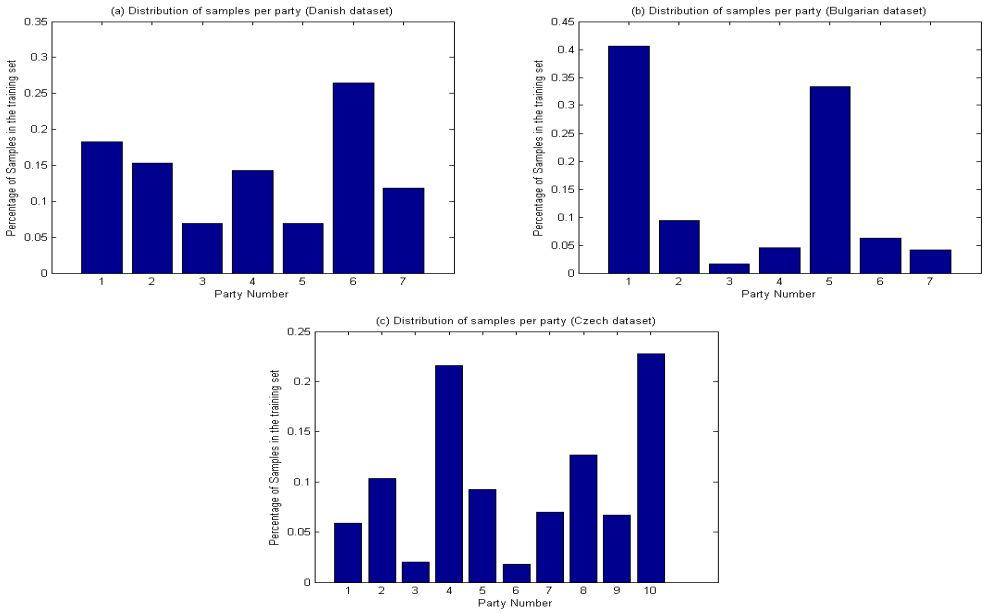


Figure 5: Distribution of samples per party in the training set for (a) Danish dataset, (b) Bulgarian dataset, (c) Czech dataset

Table 2: HMMs performance per party in the Danish dataset

Party Id	Recall	Precision	F-measure
1	0.2915	0.4925	0.3663
2	0.5103	0.4763	0.4927
3	0.7948	0.2014	0.3213
4	0.3955	0.5779	0.4696
5	0.1735	0.6101	0.2702
6	0.6132	0.7837	0.6880
7	0.5419	0.5483	0.5451

Table 3: HMMs performance per party in the Czech dataset

Party Id	Recall	Precision	F-measure
1	0.4778	0.4687	0.4732
2	0.2597	0.3262	0.2892
3	0.5411	0.2970	0.3835
4	0.6953	0.5115	0.5894
5	0.2192	0.2874	0.2487
6	0.4246	0.1836	0.2563
7	0.2889	0.7027	0.4094
8	0.4941	0.5476	0.5194
9	0.2281	0.4171	0.2949
10	0.6980	0.7183	0.7080

fifth parties have the same number of users and the smallest distribution of samples in the training set. Consequently, the HMMs for these parties achieved the worst performance exhibiting high variance between recall and prediction which reflected in low F-score. Even so, the results for the third party were better than the results for the fifth party. This shows that the users in the third party depict higher consistency on their answers and thus the HMM for this party was more effective compared to that of the fifth party.

The vote prediction performance of HMMs for the Czech dataset, shown in Table 3, varies significantly among parties. Once again the HMMs for the parties with the higher number of supporters, i.e., the tenth and fourth (see Figure 5(c)) give the best scores. The relatively low performance in vote prediction for the supporters of small parties is mainly due to insufficient number of samples. However, there are cases of parties with fewer samples, such as the third and sixth, whose HMMs performed better than parties with more samples such as the second, fifth and ninth party. By carefully examining these cases in Table 3 we see that the low number of samples reflects in unbalanced recall and precision scores, which in turn lead to low F-scores. The poor performance for the other parties is possibly due to the non-homogeneity of user profiles which leads to low scores in both recall and precision. Non-homogeneity within party supporters occurs for various reasons, such as different political background and different view for the various categories of policy statements. For instance, the supporters of the same party might have a common view on economy but totally different in EU

Table 4: HMMs performance per party in the Bulgarian dataset

Party Id	Recall	Precision	F-measure
1	0.3426	0.5114	0.4103
2	0.2832	0.3478	0.3122
3	0.1316	0.3571	0.1923
4	0.6000	0.4636	0.5231
5	0.5706	0.6884	0.6240
6	0.3409	0.1786	0.2344
7	0.3333	0.0955	0.1485

policy issues. As we explain later in the Conclusion section, within party clusters can be investigated separately by modeling data from each specific cluster through a Gaussian distribution and then generating mixture of Gaussians taking into account the ratio of each source [4, 27]. It is known that whenever the distributed data are asymmetric and multi-modal, a mixture of Gaussians can be used to model them [10].

The results for the Bulgarian dataset, shown in Table 4, are more difficult to interpret. The HMM for the first party, i.e., the one with the majority of supporters, achieved a moderate performance, while the HMM for the fourth party, which was trained on less than 100 samples, presents the second best performance. Once more, the observation made in previous datasets is confirmed: HMMs can be effectively trained even with very small training samples when these samples form a single cluster in the U-dimensional hyperspace, where U is the number of policy statements. Nevertheless, if the number of samples is adequate but there is no or low coherence between the profiles of party-supporters then the results tend to be poor.

The overall performance of HMMs in predicting vote intention in SVAAs is quite satisfactory as it can be seen in Table 5. Thus, the use of HMMs, which make use of the conditional probabilities of the VAA user answers, seems to be working. This was expected since the policy statements in VAA questionnaires are usually correlated and grouped into categories representing specific political issues. Therefore, answers to next policy statement can be 'predicted' from previous answers. Also the policy statements are answered with a specific display order, from the first to last one, and is kept constant for a specific VAA creating sequences of symbols; people who support the same party are likely to create similar sequences, since they usually share same political opinions. Thus, an HMM classifier, by utilizing answer patterns of users supporting the same party, is able to create simple and compact models that perform quite well in terms of prediction scores.

By applying HMMs to VAAs we realized that HMM classifier performance is closest to Mahalanobis classifier behavior in other VAAs, while it surpasses the performance of other machine learning algorithms, which were applied in the past to model user-party similarities (see Agathokleous *et al.* [1], Katakis *et al.* [15], Tsapatsoulis and Mendez [30], Tsapatsoulis *et al.* [29]). We noticed, however, that imperfect modeling happens either due to insufficient number of samples of a party or because of the inconsistency among users classified to the same party. Nevertheless, the non-accurate results for small parties do not critically affect the design of social recommendation, i.e., the overall vote intention predictions remains high, which is in agreement with the results reported by Tsapatsoulis *et al.* [29].

Table 5: The aggregate results of HMMs

Dataset	Recall	Precision	F-measure	MAP
Danish	0.4747	0.5647	0.5158	0.6772
Bulgarian	0.4174	0.4933	0.4522	0.6246
Czech	0.4934	0.5062	0.4997	0.6683

5. CONCLUSION

In this work we use HMM classifier in order to improve the effectiveness of social voting recommendation feature of VAAs. We based on the idea that while the users are answering the VAA policy statements they are incrementally producing sequences of observations, i.e., answer patterns, that might characterize 'typical' voters of particular parties. Thus, the ability of HMMs to capture correlations in symbol sequences would be beneficial. The performance of the proposed technique was evaluated based on the well known Recall, Precision and F-score metrics. We observed that, even if the order in which policy statements are displayed in VAAs does not actually matter, the HMMs perform very well in estimating the vote intention of users taking into account the intra-sequence correlations. This is not a surprise as the SVAAs are based on party-voters models and HMM classifier creates simple and compact models by utilizing the 'path' that users of the same party create when answering the online questionnaire. Also, the policy statements in VAAs are grouped together according to the issue category that they represent. The statements that refer on the same subject are correlated and are answered similarly by the users. Therefore, answering paths are depended and next answers depend on previous answers of same category. By finding the conditional probability in which a statement is given according to category path already occurred, the HMMs can effectively provide vote recommendation.

From our experiments we noticed that the prediction performance of HMMs depends on the consistency between the answers of the users in each party and the distribution of samples per party. Parties with the majority of users achieved the best performance in the Danish and Czech datasets. In the case of Bulgarian dataset, the HMM for the party with the highest percentage of samples presented moderate results, while the HMM for the fourth party with very few users (less than 100) achieved the second best performance. This lead us to the observation that in some cases the party-supporters profiles create a multi-modal clustering in the policy statements hyperspace (due to different political backgrounds and different views in the various categories of policy statements). In such cases the use of mixture of Gaussians [10] or different clustering techniques could be beneficial. In the near future we plan to tackle this problem by using per party and per category of policy statements HMMs. Thus, a combination of HMMs for party-supporters modeling will be pursued to account for the multi-modal distribution of VAA user profiles within the same party.

6. REFERENCES

[1] M. Agathokleous, N. Tsapatsoulis, and I. Katakis. On the quantification of missing value impact on voting

advice applications. In *Engineering Applications of Neural Networks*, pages 496–505. Springer, 2013.

[2] E. Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[3] A. Baka, L. Lia Figgou, and V. Triga. 'neither agree, nor disagree': a critical analysis of the middle answer category in voting advice applications. *International Journal of Electronic Governance*, 5:244–263, 2012.

[4] S. Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[5] C. Djouvas, K. Gemenis, and F. Mendez. Weeding out the rogues: How to identify them and why it matters for vaa-generated datasets. In *Proceedings of the 2014 European Consortium for Political Research General Conference*, pages 1–7. ECPR, 2014.

[6] C. Djouvas and N. Tsapatsoulis. A view behind the scene: Data structures and software architecture of a vaa. In *Semantic and Social Media Adaptation and Personalization (SMAP), 2014 9th International Workshop on*, pages 136–141. IEEE, 2014.

[7] K. Dyczkowski and A. Stachowiak. A recommender system with uncertainty on the example of political elections. In *Advances in Computational Intelligence*, pages 441–449. Springer, 2012.

[8] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2):81–173, 2011.

[9] J. Fivaz, J. Pianzola, and A. Ladner. More than toys: a first assessment of voting advice applications' impact on the electoral decision of voters. Technical Report 48, National Centre of Competence in Research (NCCR): Challenges to Democracy in the 21st Century, 10 2010.

[10] M. Gales and S. Young. The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304, 2008.

[11] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2000.

[12] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[13] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 135–142. ACM, 2010.

[14] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[15] I. Katakis, N. Tsapatsoulis, F. Mendez, V. Triga, and C. Djouvas. Social voting advice applications – definitions, challenges, datasets and evaluation. *Cybernetics, IEEE Transactions on*, 44(7):1039–1052, 2014.

[16] W. Khreich, E. Granger, A. Miri, and R. Sabourin. A survey of techniques for incremental learning of hmm parameters. *Information Sciences*, 197:105–130, 2012.

[17] A. Ladner, J. Fivaz, and J. Pianzola. Voting advice applications and party choice: evidence from smartvote users in switzerland. *International Journal of Electronic Governance*, 5(3-4):367–387, 2012.

[18] A. Ladner and J. Pianzola. Do voting advice applications have an effect on electoral participation and voter turnout? evidence from the 2007 swiss federal elections. In *Electronic participation*, pages 211–224. Springer, 2010.

[19] J. Lefevere and S. Walgrave. A perfect match? the impact of statement selection on voting advice applications' ability to match voters and parties. *Electoral Studies*, 36:252–262, 2014.

[20] T. Louwerse and M. Rosema. The design effects of voting advice applications: Comparing methods of calculating matches. *Acta politica*, 49(3):286–312, 2014.

[21] F. Mendez. Modelling proximity and directional logic in vaas. *Paper presented at ECPR*, 5:7, 2014.

[22] M. E. Milakovich. The internet and increased citizen participation in government. *JeDEM-eJournal of eDemocracy and Open Government*, 2(1):1–9, 2010.

[23] K. Murphy. Hidden markov model (hmm) toolbox for matlab. *online at http://www. ai. mit. edu/˜ murphyk/Software/HMM/hmm. html*, 1998.

[24] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer, 2011.

[25] M. Rosema, J. Anderson, and S. Walgrave. The design, purpose, and effects of voting advice applications. *Electoral studies*, 36:240–243, 2014.

[26] O. Ruusuvirta and M. Rosema. Do online vote selectors influence electoral participation and the direction of the vote. In *ECPR general conference*, pages 13–12, 2009.

[27] Y. Song, L. Zhang, and C. L. Giles. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, 5(4), 2011.

[28] L. Terán, A. Ladner, J. Fivaz, and S. Gerber. Using a fuzzy-based cluster algorithm for recommending candidates in e-elections. *Fuzzy Methods for Customer Relationship Management and Marketing*, 2012.

[29] N. Tsapatsoulis, M. Agathokleous, C. Djouvas, and F. Mendez. On the design of social voting recommendation applications. *International Journal on Artificial Intelligence Tools*, 24(3), 2015.

[30] N. Tsapatsoulis and F. Mendez. Social vote recommendation: Building party models using the probability to vote feedback of vaa users. In *Semantic and Social Media Adaptation and Personalization (SMAP), 2014 9th International Workshop on*, pages 124–129. IEEE, 2014.

Understanding Effects of Personalized vs. Aggregate Ratings on User Preferences

Gediminas Adomavicius
University of Minnesota
Minneapolis, MN
gedas@umn.edu

Jesse Bockstedt
Emory University
Atlanta, GA
bockstedt@emory.edu

Shawn Curley
University of Minnesota
Minneapolis, MN
curley@umn.edu

Jingjing Zhang
Indiana University
Bloomington, IN
jjzhang@indiana.edu

ABSTRACT

Prior research has shown that online recommendations have significant influence on consumers' preference ratings and their economic behavior. However, research has not examined the anchoring effects of aggregate user ratings, which are also commonly displayed in online retail settings. This research compares and contrasts the anchoring biases introduced by aggregate ratings on consumers' preferences ratings to those produced by personalized recommendations. Through multiple laboratory experiments, we show that the user preferences can be affected (i.e., distorted) by the displayed average online user ratings in a similar manner as has been shown with personalized recommendations. We further compare the magnitude of anchoring biases by personalized recommendations and aggregate ratings. Our results show that when shown separately, aggregate ratings and personalized recommendations create similar effects on user preferences. When shown together, there is no cumulative increase in the effect, and personalized recommendations tend to dominate the effect on user preferences. We also test these effects using an alternative top-N presentation format. Our results here suggest that top-N lists may be an effective presentation solution that maintains key information provided by recommendations while reducing or eliminating decision biases.

Keywords

Recommender systems; personalized ratings; aggregate ratings; preference bias; anchoring effects; laboratory experiments.

1. INTRODUCTION

Most recommender systems use user ratings for previously consumed items as inputs to the system's computational techniques to estimate preferences for items that have not yet been consumed by the individual. For example, Netflix users are asked to rate the movies they have watched on a 5-star scale (with 1 being the least liked and 5 the most liked). The Netflix recommender system then analyzes patterns of users' ratings to understand users' personal interests and predict their preferences for unseen movies. Many real-world recommender systems present these estimated user preferences in the form of system-predicted ratings to indicate expectations of how much the consumer will like items, serving as recommendations. Here we use the term "recommendation" broadly to encompass any rating that is displayed to the user with the intention to convey some item "quality" information, including the ratings at the low end of the scale (such as the recommender system's predictions that the user will dislike the item). After users experience or consume the suggested item(s), they can submit their input in the form of item ratings back to the recommender system, which are used to analyze the system's accuracy and improve future recommendations, completing a feedback loop that is central to a recommender systems' use and value (Adomavicius et al. 2013).

Recent studies show that interacting with online personalization and recommendation systems can have unintended effects on user preferences and economic behavior (Cosley et al. 2003; Adomavicius et al. 2012, 2013). In particular, users' self-reported judgments can be significantly distorted by the system's predictions. For example, Adomavicius et al. (2013) found evidence that a recommendation provided by an online system affects users' ratings for products, even immediately following consumption. Additionally, Adomavicius et al. (2012) found that personalized ratings displayed to users significantly swayed their willingness to pay for items in the direction of the system-displayed rating value.

In addition to *personalized ratings* for products (representing estimated preferences of individual users), *aggregate ratings* for products (representing population-level preference consensus) are another important type of information on which users often rely to make their product purchase or consumption decisions. While in both cases the information often is presented in an identical or nearly identical manner (e.g., as numeric scale values or star ratings), their underlying meanings are very different. However, no prior research has systematically examined or compared the potential biases caused by information presented as aggregate information about other users' evaluations vs. personalized preference predictions from a recommender system. Therefore, we explore four issues related to the impact of aggregate vs. personalized ratings:

(1) The *bias* issue: Are users' self-reported preference ratings for products drawn toward the displayed aggregate values? In other words, do preference biases that have been observed with personalized ratings extend to non-personalized, aggregate ratings?

(2) The *relative effect size* issue: Observing which type of information results in a higher effect size: aggregate or personalized ratings?

(3) The *combination* issue: What is the combined effect of providing both aggregate and personalized ratings, as compared to receiving one alone?

For the fourth issue, we first note that, instead of displaying rating information as numeric values (e.g., personalized system-predicted ratings or mean aggregate ratings), sometimes systems use rating information to compile and display top-N item lists of "best" (or recommended) items. As examples, many news websites recommend the top-10 articles to their readers based on their interests and browsing histories, and Amazon suggests lists of products that their customers might find interesting. These

recommendations are merely displayed on the webpage as a list of items (either ranked or not-ranked) often with no underlying rating values. Studying this type of presentation format provides a robustness check on the biasing effects of personalized and aggregate ratings.

(4) The *presentation format* issue: Are users' self-reported preference ratings for products still influenced by the displayed information, when the aggregate and personalized ratings are used to produce the recommendations as a list of the top-N items but without displaying explicit numeric rating values?

We conducted three controlled laboratory experiments, in which the recommendations based on aggregate and/or personalized ratings presented to participants were manipulated to answer the aforementioned research questions. In all studies, participants were asked to read a number of jokes, reporting their preference rating immediately after reading each joke.

2. BACKGROUND

Research studies in decision making, behavioral economics, and marketing have consistently observed judgmental biases across a variety of context settings (e.g., see contributions in Gilovich et al. 2002). In lay usage, the term "bias" has a negative connotation, suggesting a negative prejudice. In behavioral economic and decision research, however, the word bias is used in a more agnostic manner to represent a systematic pattern of deviation from a norm or rational standard of judgment (e.g., Haselton et al. 2005). Decision biases are not always detrimental for the decision maker; instead, the term bias highlights predictable tendencies that judgments follow under certain decision conditions.

In the context of users responding to recommendations, we define bias relative to a rational standard that is at least implicit, if not explicit, in all real-world instances of their use. The presumption is that the consumer's stated preference rating is a non-adulterated expression of their preference for the product or experience itself, as tailored to the provided scale. This should particularly be true when there is no delay between the experience or consumption of the item being rated and the reporting of the preference. This standard parallels the normative principle of invariance described and tested by Tversky and Kahneman (1986) in their discussion of framing effects as a judgmental bias. If a personalized recommendation significantly impacts the stated preference, signaling an adulteration of the preference, then we say that a decision bias has been introduced by the recommendation.

In identifying bias in this way, it is important to recognize the timing of the connection between the system's recommendation (which may include personalized or aggregate ratings for an item) and the consumer's subsequently submitted preference rating. Prior to experiencing the item, consumers seek and receive recommendations as a way of guiding their choices, purchases, and/or expectations concerning the item. At this *pre-consumption stage*, the recommendations represent a highly valuable service to consumers, providing help for finding and selecting relevant items and managing the potential information overload in many online settings. Once the item has been experienced, however, the recommendation is not assumed to provide value in the assessment of the user's inherent preference for that item. This is particularly true if the item has just been consumed, i.e., there is no potential uncertainty about the experience due to recall effects (e.g., when users are trying to remember how they felt about a movie that they saw a year ago). Recommendations are designed to provide value at the pre-consumption stage; if they impact *post-consumption* preferences as well, this represents a bias relative to the presumed standard of unpolluted preference.

Within the context of recommender systems, a few studies have explored how personalized system-predicted ratings influence online consumer behavior. These studies have shown strong and consistent evidence that ratings provided by consumers are biased toward system-generated recommendations when consumers construct their judgments for products. For example, Cosley et al. (2003) explored the effects of system-generated recommendations on user re-ratings of movies and found that users showed high test-retest consistency when no prediction was provided. However, when users re-rated a movie while being shown a "predicted" personalized value that was altered upward or downward from their system's actual prediction by a single fixed amount of one point (i.e., providing a higher or lower prediction), users tended to give higher or lower ratings, respectively, as compared to a control group receiving the system's actual predictions. This showed that system predictions could affect users' ratings based on preference recall, for movies seen in the past and now being evaluated. Additional recall-based effects are explored by Bollen et al. (2012).

More recently, Adomavicius et al. (2013) examined system effects in three laboratory studies for items elicited at the time of item consumption. The design removes possible explanations deriving from the preference uncertainty that can be present at the point of recall, i.e., when trying to evaluate one's preferences for an item that may have been experienced long ago. In this setting, one's preferences should arguably be based solely on the immediate experience of the item; no uncertainty is present. Even without a delay between consumption and elicited preference, consumers' preference ratings were consistently influenced by the system-generated personalized recommendations. The effect was observed across different content domains (TV shows and jokes). And, the effect obtained whether the recommendation was seen before or after watching a TV show; so, an explanation based on priming the viewers' expectation for the upcoming experience was not supported. Consistently, the displayed system predictions, when perturbed to be higher or lower, affected the submitted consumer ratings to move in the same direction.

Further, recent research has found that the system-generated ratings can significantly affect consumers' economic behavior with respect to the suggested items (Adomavicius et al. 2012). Using three controlled experiments in the context of digital song purchases, the authors found strong evidence that song recommendations substantially affected participants' willingness to pay for the songs, even when controlling for participants' preferences and demographics. The effects persisted even when item uncertainty was reduced, in this case by forcing participants to listen to song samples prior to pricing the songs. The effect also persisted when scale compatibility issues were removed. Scale compatibility is another common explanation for biases whereby using the same scale for predictions and user ratings creates a demand effect to increase the correspondence between stimulus and response. In the study, willingness-to-pay judgments were expressed using a 0-99 scale, i.e., in U.S. cents, and system ratings where expressed using a typical 1-5 star scale. Thus, the effect of system recommendations is not purely an effect of reacting to a numerical value on a common scale. Overall, the biases resulting from system recommendations on preference judgments have been shown to be robust across a variety of digital goods, settings, and conditions.

Further, these biases can be potentially harmful in several ways (Cosley et al. 2003; Adomavicius et al. 2013). From the consumers' perspective, recommendation biases can distort (or

manipulate) their preferences and their purchasing behavior and, therefore, lead to distortions in their self-reported preference ratings and suboptimal product choices. From the firm's perspective (e.g., Amazon, Netflix), these biases may allow third-party agents to manipulate the recommender system so that it operates in their favor. This would reduce consumers' trust in the recommender system and harm its value in the long term. From the system designers' perspective, the distorted user preference ratings that are subsequently submitted as consumers' feedback can pollute the inputs of the recommender system, reducing its effectiveness.

In addition to personalized system-predicted ratings, *aggregate ratings* represent an alternate source of information about item "quality" that is directly relevant to users' decision making. In particular, they can be viewed as non-personalized (i.e., same for all users) recommendations that indicate the population-level consensus about the general quality level of a given item. This information may be derived from peer ratings or from aggregating sales, download, or click data. Similar to personalized recommender system predictions, aggregate rating information may be communicated to consumers in the form of numeric values, such as mean peer ratings, or can be used to construct top-N lists of generally best-liked items.

Note that the nature of the rating effects can be studied from both *macro-level* and *micro-level* perspectives. The market, macro-level perspective investigates the market effects of how ratings and recommendations impact sales, downloads, or other aggregate outcomes of interest to retailers. From the consumer, micro-level perspective, the interest is on how ratings and recommendations impact the behaviors of individual users.

The study of macro-level outcomes has been an active area of research investigation in recent years. For example, with respect to the effects of providing aggregate ratings, Tucker and Zhang (2011) studied the impact of popular bestseller listings based on previous clicks upon the number of future clicks received. Similarly, Godinho de Matos et al. (2016) investigated the influence of peer ratings, expressed as a list of most popular movies, on market sales within a natural field experiment. Also, using an experimental methodology by creating a music market of unknown songs and artists, Salganik et al. (2006) manipulated whether or not the participants saw the number of downloads made by others and studied the effect of this social influence on market factors. An example of academic research that investigated the market-level effects of personalized recommendations is a study by Fleder and Hosanagar (2009) who, using analytical modeling and simulation, suggested that recommendation systems can lead to a rich-get-richer effect for popular products, resulting in a decrease in sales diversity in the aggregate. Somewhat in contrast, results of Fleder and Hosanagar also suggested that personalization technologies help users to widen their interests, increasing the likelihood of commonality with others.

In contrast, the micro-level effects have been underexplored in research literature, especially with respect to the aggregate ratings and their impact on individual consumer preferences. Therefore, we focus on this issue in our current study: How do personalized system-predicted ratings and aggregate peer ratings compare and contrast as influences on individual users' reactions, particularly in the preference bias that they produce?

This comparison is particularly interesting, because the influences of personalized vs. aggregate ratings on user behavior are hypothesized along quite different psychological mechanisms. For example, supplying aggregate data within a music market, Salganik and Watts (2008) demonstrated the effect that aggregated popularity feedback had upon individual-level responses in terms of choices to listen to and download songs. The mechanism for the effect derives from social motivations, grounded in the literatures on social influence. The general dynamic is one in which the consumer engages in a form of observational learning of how to behave based on the behavior of others. In contrast, personalized recommendations do not arise from social comparison. Depending on the recommendation algorithm, the personalized system-predicted rating may or may not have any connections with others' behavior. For example, content-based algorithms depend on matching feature characteristics, not on the preferences of other users (Ricci et al. 2011). Even algorithms that incorporate preferences of other users, e.g., collaborative filtering techniques (Ricci et al. 2011), generally do not make the connection explicit or obvious to the consumer. Therefore, rather than mechanisms grounded in social psychology, the effects of personalized ratings can be posited on bases of anchoring, information integration, and processing explanations. For example, one proposed mechanism is in terms of scale compatibility, as mentioned above. Another sample mechanism proposed for the effects of personalized recommendations is an information integration explanation whereby the system-predicted rating is perceived as a piece of information that the user should use in constructing their judgment (cf. Mussweiler and Strack 1999).

3. STUDY 1: Individual Effects of Aggregate vs. Personalized Ratings

3.1 Design

This study focused on research questions (1) and (2). All the studies described in this paper involved the consumption and rating of jokes, so the participant population required no special characteristics. Participants were 118 recruits from a US college's research participant pool. Participants were paid a fixed $10 fee for completing the study. Demographic features of the sample are summarized in Table 1 separately for each of the two conditions of the between-subjects component of the design. Participant characteristics are comparable between the two treatment groups. The mean time for completing the study was 29.03 minutes, which suggests subjects invested ample time and that fatigue was not an issue.

Table 1. Demographic characteristics of participants in Study 1.

	Personalized Rating	**Aggregate Rating**
# of Participants	59	59
% Female	45.8%	47.5%
Age: Mean (SD)	23.6 (8.70)	24.0 (9.03)
% Native English Speaker	50.9%	61.0%
% Undergraduate	64.5%	55.7%

Our study used 100 jokes from the Jester joke database, which has been extensively used in prior literature (Goldberg et al. 2001; Adomavicius et al. 2013). Jokes are stimuli that can be experienced in the lab session, so that the readers' preference ratings can be gathered immediately after the reading of each joke; there is no uncertainty of preference due to memory effects. As noted in Section 2, the standard assumption in such a situation is that the user's rating should provide an unadulterated expression of the reader's preference, forming the normative expectation against which bias is defined.

The between-subjects manipulation in the study is based on the type of rating information that is presented to study participants: personalized vs. aggregate ratings. In other words, the information

is presented as either a personalized rating from a recommender system or as a mean rating of other users. Each participant saw only one type of rating information, either the aggregate or personalized ratings, for all the jokes.

Using a 5-star rating scale (allowing half-star ratings), participants first evaluated 50 jokes, which were randomly selected from the list of 100 and randomly ordered. These ratings provided a guise of collecting data from which to derive personalized recommendations, and also allowed us to calculate rating predictions for use in the analysis as a control for individual differences in preference.

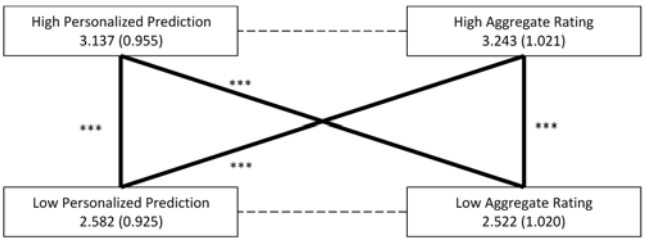
Next, the subjects received 45 jokes with rating-based information displayed. Half of the subjects randomly received the information in the form of aggregate ratings displayed as "Average user rating of this joke is: X (out of 5)", while the other half in the form of personalized ratings displayed as "Our system thinks you would rate the joke as: X (out of 5)". Here X is a specific rating value that was assigned separately for each joke. In both of these treatment groups (referred to as AggregateOnly and PersonalizedOnly groups), the participants saw 45 jokes in three within-subjects conditions. Specifically, 20 of these jokes were assigned to the High condition, which consisted of randomly-generated high values between 3.5 and 4.5 stars (drawn from a uniform distribution); another 20 jokes were assigned to the Low condition, which consisted of randomly-generated low values between 1.5 and 2.5 stars (drawn from a uniform distribution); and the remaining 5 jokes were assigned as the Medium condition which included randomly generated values between 2.5 and 3.5 (drawn from a uniform distribution). These 45 jokes were randomly intermixed. The Low and High conditions were oversampled since the High-Low comparison is the test of bias in this setting – i.e., whether the participants would report their post-consumption preference rating differently after being exposed to High vs. Low rating from the system. The Medium condition is included so that the presented ratings could cover the entire spectrum of the 1-5 rating scale; this helps to avoid possible credibility issues caused by bipolar recommendations (i.e., having either very high or very low ratings displayed). The responses to the Medium rating items are only useful in addressing the more peripheral issue of whether there is asymmetry in the bias between the High and Low ranges (comparing the difference of differences between High-Medium and Medium-Low), an issue that is not addressed by this paper. Hence, the Medium ratings are not analyzed here.

Finally, participants completed a short survey that collected demographic and some other individual information for use in the analyses (e.g., see Table 2 for demographics).

3.2 Results

The pairwise t-test comparisons for the High and Low treatments are illustrated in Figure 1. Both aggregate and personalized ratings generate substantial effects on post-consumption preference ratings. Users reported significantly inflated preference ratings in the high conditions as compared to the low conditions (by more than a half-star on average, overall). This result for the personalized ratings is consistent with prior results (e.g., Adomavicius et al. 2013). The result for the aggregate ratings addresses research issue (1): Even though the two types of recommendation represent very different information, they both tend to generate biases of preference ratings.

With respect to research issue (2), we note that the two Low conditions (Low personalized ratings vs. Low aggregate ratings) do not significantly differ, and that the same holds for the two High conditions. This provides evidence that aggregate and personalized ratings, when presented individually, generate similar levels of preference bias.



Note: ***p < .001, ** p < .01, [----] p > .05

Figure 1. Mean (standard deviation) of self-reported user preference ratings after observing either High vs. Low personalized or aggregate ratings. (All tests are one-tailed except for the tests represented by the horizontal lines in the figure. Unlike the others, these tests have no prior hypothesized direction, so two-tailed tests are performed.)

To get a more direct answer for research issue (2), with controls for possible confounding factors, we also employed regression analysis. Specifically, the repeated-measures design of the experiment, wherein each participant was exposed to both high and low ratings in a random fashion, allows us to model the aggregate relationship between shown ratings (either personalized or aggregate) and user's submitted post-consumption preference ratings while controlling for individual participant differences. While we do not include the full regression analysis here due to the space limitations, the results confirm what we have observed in Figure 1 – i.e., the social, population-level information provided by aggregate ratings creates a level of bias comparable to that produced by personalized recommendations despite their different underlying mechanisms.

4. STUDY 2: Combined Effects of Both Personalized and Aggregate Ratings

4.1 Design

We recognize that, in some instances, retailers provide both pieces of information (personalized *and* aggregate ratings) as aids to the consumer. Study 2 addresses research issue (3), i.e., the issue of the *combined* effects of these two different types of ratings.

A US college's research participant pool provided 55 participants who were paid a fixed $10 fee for completing the study. None of the participants from Study 1 were allowed to enroll in Study 2. Demographic features of the sample are summarized in Table 2 for the two groups (discussed below) of the between-subjects component of the design. Participant characteristics are comparable between the two treatment groups and to those in Study 1 drawn from the same population. The mean time for completing the study was 29.97 minutes.

Table 2. Demographic characteristics of participants in Study 2.

	Personalized Rating First	**Aggregate Rating First**
# of Participants	28	27
% Female	35.7%	48.2%
Age: Mean (SD)	24.1 (7.07)	25.7 (12.26)
% Native English Speaker	67.9%	70.4%
% Undergraduate	55.4%	56.78%

The objective of Study 2 is to examine the relative importance of personalized vs. aggregate ratings when both types of information are displayed, controlling for any order effects. Participants were

randomly assigned into one of two treatment groups. Participants in both groups received both personalized and aggregate ratings for each joke displayed to them. The first group received the personalized rating first, followed by the aggregate rating (i.e., the PersonalizedFirst group). The second group saw the ratings for each joke in the reverse order (i.e., the AggregateFirst group).

The study followed a similar procedure as Study 1. Participants first read 50 randomly selected jokes from a database of 100, being asked to provide their preference for each joke using a 5-point rating scale. Each participant was then asked to rate 45 additional jokes along with both personalized and aggregate ratings displayed.

In both treatment groups, the 45 jokes were randomly assigned into five within-subjects conditions. 40 of the jokes occupied a 2×2 within-subjects design crossing High and Low values of personalized and aggregate ratings. 10 jokes were assigned to the HighP-HighA condition that consisted of high values for both personalized and aggregate ratings, 10 jokes to the LowP-LowA condition that consisted of low values for both personalized and aggregate ratings; 10 jokes to the LowP-HighA condition that consisted of low personalized and high aggregate ratings, and 10 jokes to the HighP-LowA condition that consisted of high personalized and low aggregate ratings. Similar to Study 1, all the high values are randomly generated values between 3.5 and 4.5 stars, and all the low values are randomly generated values between 1.5 and 2.5 stars, drawn from a uniform distribution. The remaining 5 jokes were assigned to the Medium condition, which included randomly generated values between 2.5 and 3.5 for both personalized and aggregate ratings. As in Study 1, the Medium condition was included simply to have a credible representation of ratings from the entire spectrum of the 1-5 rating scale; the Medium ratings are not used in any subsequent analyses in the paper. The 45 jokes were randomly intermixed.

Finally, the participants were asked to complete a survey about their demographic information and joke preferences.

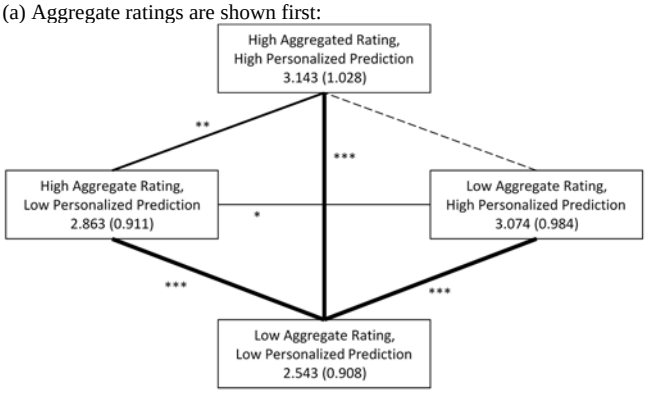
4.2 Results

Figure 2 illustrates the mean self-reported preference ratings submitted by Study 2 participants, for the cases when aggregate ratings are shown first (Figure 2a) and when personalized ratings are shown first (Figure 2b). Starting with the vertical line in each diagram, we see a strongly significant difference between when both personalized and aggregate ratings are high vs. when both ratings are low, indicating a clear bias effect in both cases, consistent with Study 1. The horizontal line in each figure indicates a stronger impact of personalized ratings compared to aggregate ratings when both appear together and signal in opposite directions (one with a high recommendation and the other low). Thus, the pairwise contrasts suggest that, when both types of ratings are present, personalized recommendations seem to be taken into account by users more strongly (and, hence, generate more significant biases) than aggregate ratings regardless of the presentation order.

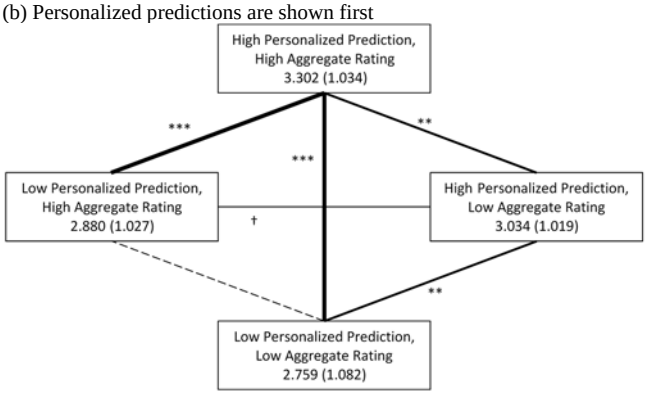
This pattern is also supported by the diagonal lines in Figure 2. Beginning with the negatively sloped diagonals in the figure, i.e., when the aggregate rating goes from low to high, holding the valence of the personalized prediction fixed, the effect is variable. In each case, one of the comparisons is statistically significant, and the other is not. Specifically, a low to high difference for aggregate ratings is only observed when the aggregate ratings are preceded by high personalized predictions (Figure 2b) or followed by low personalized predictions (Figure 2a). In contrast, from the comparisons indicated by the positively sloped diagonals in the figure, i.e., when the personalized prediction goes from low to high, holding the valence of the aggregate ratings fixed, we see a clear, consistent, statistically significant biasing effect.

By comparing Figures 1 and 2, one can observe that the High vs. Low effect magnitudes are of around 0.55-0.72 when ratings are displayed individually, and around 0.55-0.6 when two ratings are displayed together. In other words, the preliminary evidence suggests that the cumulative effect of both ratings is not greater than the effect of either one of the ratings individually.

(a) Aggregate ratings are shown first:



(b) Personalized predictions are shown first



Note: ***p < .001, ** p < .01, * p < .05, † p < .10, [----] p > .10

Figure 2. Mean (standard deviation) of user preference ratings when both personalized and aggregate ratings were displayed. (All tests are one-tailed except for the tests represented by the horizontal lines in the figure. Unlike the others, these tests have no prior hypothesized direction, so two-tailed tests are performed.)

To test for the robustness of the combined effects of the two types of ratings (as compared to when each rating is shown individually) and to allow us to control for possible confounding factors, we conduct regression analyses. To do so, we pool the conditions from Studies 1 and 2 (that sampled from the same participant population) to compare the effect sizes when different amounts of information are presented to users. Recall that the High-Low comparison is the test of bias in our setting. Therefore, we include all the observations from Study 1, while for Study 2 we only consider the cases when both high personalized and high aggregate ratings (i.e., HighP-HighA) and when both low personalized and low aggregate ratings (i.e., LowP-LowA) are displayed to users. The random-effects GLS model using robust standard errors, clustered by participant, and using participant-level controls represents our model for the analysis:

$$UserRating_{ij} = b_0 + b_1(High_{ij}) + b_2(Group_i) + b_3(High_{ij}*Group_i) + b_4(Deviation_{ij}) + b_5(PredictedRating_{ij}) + b_6(AdditionalControls_{ij}) + u_i + \varepsilon_{ij}$$

In the model, $UserRating_{ij}$ is the submitted rating for participant *i* on joke *j*. $High_{ij}$ is a binary variable that indicates whether the shown rating for participant *i* on joke *j* is a high or low artificial rating. Thus, the coefficient on $High_{ij}$ measures the difference in user preference ratings between high and low conditions (manipulated within-subjects), which is our operationalization of bias. To the extent that users are influenced by the observed information, their submitted preference ratings will be shifted up when seeing High ratings and shifted down when seeing Low ratings. Thus, the High/Low difference is an indicator of the bias created by the observed information. $AggregateDev_{ij}$ and $PersonalizedDev_{ij}$ are derived variables that capture the deviations between the actually shown rating (aggregate and personalized, respectively) for participant *i* on joke *j* and the expected value for the shown ratings in the corresponding condition. Recall that the two ratings seen by the participant were manipulated independently to introduce randomness into the values for the high and low rating conditions by drawing from uniform distributions: High [3.5, 4.5] and Low [1.5, 2.5]. Thus, the deviation variables are computed by either subtracting 4.0 from the shown rating (High condition) or 2.0 from the shown rating (Low condition). $Group_i$ denotes different between-subject conditions with different information displays. For Model 1, the groups are *AggregateOnly* (from Study 1) and the two order conditions from Study 2: *AggregateFirst* and *PersonalizedFirst*. For Model 2, the AggregateOnly condition is replaced by the *PersonalizedOnly* condition. We also include the interaction term between $Group_i$ with the rating value (i.e., $High_{ij}$*$Group_i$). The interaction term examines whether the effect size of showing high vs. low information differs among groups. $PredictedRating_{ij}$ is the system-predicted rating for participant *i* on joke *j*, calculated after the fact.[1] The inclusion of this term provides an important control for differing expected joke preferences among participants. The collection of the ratings on the first 50 jokes in the procedure (without recommendations) allowed us to calculate these predicted ratings for all subjects. $AdditionalControls_{ij}$ is a vector of joke and participant-related variables. The controls included in the model were: the joke's funniness (average joke rating in the Jester dataset, continuous between 0 and 5), participant age (integer), gender (binary), school level (undergrad yes/no binary), whether they are native speakers of English (yes/no binary), whether they thought recommendations in the study were accurate (interval five-point scale), whether they thought the recommendations in the study were useful (interval five-point scale), and whether they thought that recommendations in general were useful (interval five-point scale).

The study utilized a repeated-measures design with a balanced number of observations on each participant. To control for participant-level heterogeneity, the composite error term ($u_i + \varepsilon_{ij}$) includes the individual participant effect u_i and the standard disturbance term ε_{ij}. A random-effects model is used for participant heterogeneity, since these individual-specific effects are uncorrelated with the randomly applied treatment conditions.

Table 3 summarizes our pooled regression analyses. The models are analyzed separately to allow us to use a full set of control variables for each regression model. Model 1 includes the conditions in which aggregate ratings are displayed, i.e., AggregateOnly from Study 1, and the HighP-HighA and LowP-LowA conditions from Study 2. Model 2 includes the conditions in which personalized predictions are displayed, i.e., PersonalizedOnly from Study 1 and the HighP-HighA and LowP-LowA conditions from Study 2.

Interestingly, in both models, the interaction term $High_{ij}$*$Group_i$ is statistically insignificant. There is no evidence that the effect size measured in high-low difference differs among groups with different information displays. In other words, when aggregate and personalized ratings are presented together (as explored in Study 2) regardless of their order, there is no evidence that they increase preference biases compared to displaying either aggregate or personalized recommendations alone. In other words, our results suggest a substitutionary effect between personalized and aggregate ratings on users' self-reported preferences (so that the cumulative effect is no greater than the effect of only one).

Table 3. Pooled Regression analyses for Studies 1 and 2.

DV: UserRating	Model 1: AggrOnly (Study1), All Data (Study2)	Model 2: PersOnly (Study1), All Data (Study2)
	Coefficient (SE)	Coefficient (SE)
High	0.718(0.06)***	0.551(0.068)***
Group		
(Baseline: AggregateOnly or PersonalizedOnly)		
AggregateFirst	-0.037(0.1)	-0.06(0.091)
PersonalizedFirst	0.118(0.103)	0.109(0.097)
High * Group		
High * AggregateFirst	-0.078(0.138)	0.074(0.142)
High * PersonalizedFirst	-0.135(0.138)	0.018(0.142)
AggregateDev	0.249(0.048)***	
PersonalizedDev		0.264(0.055)***
PredictedRating	0.838(0.073)***	0.679(0.054)***
Control		
jokeFunniness	0.163(0.105)	0.182(0.088)*
Age	-0.003(0.004)	0(0.002)
Male	0.006(0.067)	0.045(0.048)
Undergrad	0.09(0.077)	0.025(0.055)
Native	0.001(0.065)	-0.133(0.046)*
AggregateAccurate	-0.032(0.028)	
AggregateUseful	0.009(0.022)	
GeneralAggregateUseful	0.011(0.014)	
PersonalizedAccurate		-0.07(0.025)*
PersonalizedUseful		0.067(0.021)***
GeneralPersonalizedUseful		-0.002(0.015)
Constant	-0.368(0.301)	0.107(0.31)
N	114	114
R^2 within-subject	.2204	.1696
R^2 between-subject	.6148	.6605
R^2 overall	.3024	.2474
χ^2	992 (15 df), p < .0001	676 (15 df), p < .0001

5. STUDY 3: Recommendation List Effects

5.1 Design

In Studies 1 and 2, the recommendations were presented to users as values along a 1-5 scale. As noted earlier, another common format for providing recommendations is in the form of lists of recommended items. The numeric values corresponding to the list items often are not displayed, even when these recommendations are generated by selecting items based on the personalized rating

[1] We applied the well-known item-based collaborative filtering (CF) technique (Deshpande and Karypis 2004; Sarwar et al. 2001) to implement a recommender system that estimated users' preference ratings for the jokes. Item-based CF is one of the most popular techniques used in real-world applications because of its efficiency and accuracy. This technique allows us to precompute the main portion of our recommendation model (i.e., the similarity scores between items based on their rating patterns) in advance based on the extensive Jester rating dataset.

predictions or aggregate peer rating values. In Study 3, we examine whether the recommendations (based on either aggregate or personalized ratings) presented as a list of top-N items can lead to bias in users' reported preference ratings. Doing so explores the generalizability of the results of Studies 1 and 2, addressing research issue (4): As a non-numerical format (derived from numerical information), when compared to explicitly provided numeric ratings, do top-N recommendation lists create bias in users' self-reported post-consumption preference ratings?

Recruited from the same population as for Studies 1 and 2, 184 new participants were paid a fixed $10 fee for completing the study. Table 4 shows demographic features of the sample across all conditions of the between-subjects component of the design. Participant characteristics are comparable to those of Studies 1 and 2. The mean time for completing the study was 20.13 minutes.

Table 4. Demographic characteristics of participants in Study 3.

# of Participants	184
% Female	58.7%
Age: Mean (SD)	22.8 (7.70)
% Native English Speaker	76.6%
% Undergraduate	77.7%

The procedure followed that used in Studies 1 and 2. Participants first went through a list of 50 randomly selected jokes from the 100 jokes in the database, providing ratings using the 5-point rating scale. Participants then saw 20 additional jokes displayed as a list, and they rated the jokes using the same 5-point rating scale. Finally, the participants were asked to complete a survey about demographic information and joke preferences.

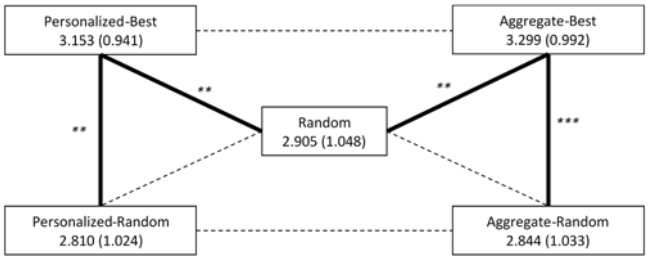
Table 5. Experimental conditions used in Study 3.

Group	N	System Description	Actual Operationalization
Random	34	"These are 20 additional jokes from our database (in no particular order)."	A list of 20 randomly-selected jokes were displayed to all participants.
Personalized-Random	34	"Based on the ratings you provided in the previous step, our recommender system has made predictions of your personal preferences on the remaining jokes in our database. These are 20 most recommended jokes for you (in no particular order)."	A list of 20 randomly-selected jokes were displayed to all participants.
Personalized-Best	39		A list of top 20 jokes with the highest predicted user preference ratings were selected for each participant; each participant saw a different list of jokes.
Aggregate-Random	34	"Based on other users' ratings, we have computed average funniness for all jokes in our database. These are 20 most overall liked jokes (in no particular order)."	A list of 20 randomly-selected jokes were displayed to all participants.
Aggregate-Best	43		A list of top 20 jokes with the highest mean user ratings were selected.

Each participant was randomly assigned into one of five treatment groups: Random, Personalized-Random, Personalized-Best, Aggregate-Random, and Aggregate-Best. Table 5 summarizes the five conditions and the number of respondents in each condition. In the Random condition, participants were told that the additional 20 jokes were randomly selected from an existing database and displayed in no particular order, and these jokes were indeed random selections from the database in the actual operationalization. For the two personalized conditions, both groups of participants were told that the item list contains the top jokes selected by a recommender system based on their preferences. The difference between Personalized-Random and Personalized-Best lies in the actual operationalization of selecting the jokes. In the Personalized-Random condition, the jokes displayed to the participants were actually randomly selected jokes from the database. In the Personalized-Best condition, the jokes were the *actual* top 20 jokes that had the highest predicted ratings estimated by the well-known item-based collaborative filtering recommendation algorithm (Deshpande and Karypis 2004; Sarwar et al. 2001). Similarly, for the two aggregate conditions, both groups of participants were told that the item list contains the top jokes selected according to aggregate user ratings on these jokes. In the Aggregate-Random condition, the jokes displayed to the participants were actually randomly selected jokes from the database. In the Aggregate-Best condition, the jokes were the actual top 20 jokes that had the highest mean user ratings based on the Jester dataset. To control for joke funniness, we displayed the same list of 20 jokes to all participants in the Personalized-Random, Aggregate-Random and Random conditions, albeit the display order was shuffled for each person.

5.2 Results

Figure 3 illustrates the pairwise t-tests that compare mean user submitted ratings for the experimental conditions. When random jokes were provided, identifying the jokes as recommended, either based on system predictions (Personalized-Random) or aggregate ratings (Aggregate-Random), did not lead to significantly higher user ratings compared with the Random group. In other words, the top-N information presentation format did not introduce bias in users' submitted ratings. Figure 3 also suggests that the consumers were not just generally insensitive. On average, participants who received actual top-N lists, either based on personalized (Personalized-Best) or aggregate ratings (Aggregate-Best), provided significantly higher ratings on these items than participants who received random recommendations. Note that, in this case, such rating differences are not necessarily indicative of bias in user preferences, since the jokes displayed in the Personalized-Best and Aggregate-Best conditions were likely better jokes for the participants.



Note: ***p < .001, ** p < .01, [----] p > .05

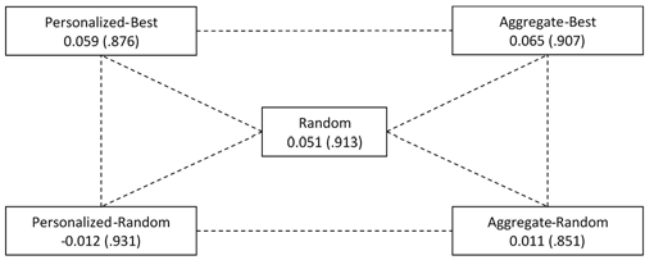
Figure 3. Contrasts of mean preference ratings by experimental condition. (All tests are two-tailed.)

In order to separate item quality from potential bias, we conducted two further analyses. First, we adjusted the preference ratings for the viewed jokes using a rating drift measure (Adomavicius et al. 2013) defined as:

Rating Drift = Submitted Rating – System-Predicted Rating.

The system- predicted rating represents the rating of a user-joke combination as predicted by the recommendation algorithm (an item-based collaborative filtering method, in our case). Submitted rating is the user's reported rating after reading the joke. So, positive/negative rating drift values represent situations where the user's submitted rating was higher/lower than the system-predicted rating. In that the predicted rating captures a valid indicator of user's preferences based on their initial joke responses (which has been demonstrated in prior work, e.g., by Adomavicius et al. (2013)), rating drift is a measure that removes a component of

individual preference from the user rating, leaving a measure that is more representative of possible bias (though still not pure, since the predicted ratings are not 100% accurate). Still, the contrast t-tests of rating drift across treatments in Study 3 are highly illuminative, as illustrated by Figure 4. As observed, rating drifts for all five experimental groups had small values that ranged from -0.012 to 0.065, suggesting that user-submitted ratings were statistically indistinguishable from system-predicted ratings. Also, the differences in rating drifts between Random and any treatment group were insignificant, suggesting that item lists presented in different ways (i.e., based on personalized recommendations, aggregate user ratings, or random lists) did not pull users away from their preferences as captured by the system-predicted ratings.



Note: [----] p > .05

Figure 4. Contrasts of mean rating drifts by experimental condition. (All tests are two-tailed.)

6. CONCLUSIONS

Prior research has shown that system recommendations can impact users' self-reported preference ratings, which can have deleterious effects (Cosley et al. 2003; Adomavicius et al. 2013). We extend this stream of research to investigate the decision biases introduced by aggregate peer ratings on users' post-consumption preference ratings. Through laboratory experiments, we first demonstrate that the self-reported preference rating (for a specific consumed item) can be strongly biased not only by observing personalized, system-predicted ratings, but also non-personalized, aggregate user ratings. We further demonstrate that, when personalized and aggregate ratings are displayed together, there is no cumulative increase in effect and that users tend to focus more attention on personalized ratings. Finally, we show that alternative recommendation displays that use top-N lists instead of individual item rating information seem to greatly reduce, if not remove, the biases observed in prior studies – in other words, this format appears to be a promising alternative for recommending items to users without introducing decision biases. This may be because top-N lists do not include explicit values (i.e., the individual system-predicted ratings or aggregate user ratings), which are likely causing the biases observed in prior studies. These results provide several obvious practical implications for the design of recommender system and online retail interfaces and displays.

7. REFERENCES

[1] Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2012. Effects of Online Recommendations on Consumers' Willingness to Pay. *ACM RecSys 2012 Workshop on Human Decision Making in Recommender Systems*, Dublin, Ireland.

[2] Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* (24:4), pp. 956-975.

[3] Bollen, D., Graus, M., and Willemsen, M.C. 2012. Remembering the Stars? Effect of Time on Preference Retrieval from Memory. In *Proceedings of the Sixth ACM Conference on Recommender Systems* (RecSys 2012). ACM, New York, NY, USA, 217-220.

[4] Cosley, D., Lam, S., Albert, I., Konstan, J.A., and Riedl, J. 2003. Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI 2003), Fort Lauderdale, FL, pp. 585-592.

[5] Deshpande, M., and Karypis, G. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* (22:1), pp. 143-177.

[6] Fleder, D., and Hosanagar, K. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* (55:5), pp. 697-712.

[7] Gilovich, T., Griffin, D., and Kahneman, D. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment* (1 ed.), Cambridge University Press.

[8] Godinho de Matos M., Ferreira P., Smith M.D. and Telang R. 2016. Culling the herd: Using real-world randomized experiments to measure social bias with known costly goods. *Management Science*.

[9] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. 2001. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval* (4:2), pp. 133-151.

[10] Haselton, M.G., Nettle, D., and Andrews, P.W. 2005. The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology*, D.M. Buss (ed.), Hoboken NJ: John Wiley & Sons, pp. 724–746.

[11] Mussweiler, T., and Strack, F. 1999. Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. *Journal of Experimental Social Psychology* (35:2), 3//, pp. 136-164.

[12] Ricci, F., Rokach, L., Shapira, B., and Kantor, P. (eds.) *Recommender Systems Handbook*, Springer, 2011.

[13] Salganik, M.J., Dodds, P.S., and Watts, D.J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* (311:5762), pp. 854-856.

[14] Salganik, M.J., and Watts, D.J. 2008. Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market. *Social Psychology Quarterly* (74:4), Fall, p. 338.

[15] Sarwar, B., Karypis, G., Konstan, J.A., and Riedl, J. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. *10th International WWW Conference*, Hong Kong, pp. 285-295.

[16] Tucker, C., and Zhang, J. 2011. How Does Popularity Information Affect Choices? A Field Experiment. *Management Science* (57:5), pp. 828-842.

[17] Tversky, A., and Kahneman, D. 1986. Rational Choice and the Framing of Decisions. In *Rational Choice: The Contrast between Economics and Psychology*, R.M.Hogarth and M.W. Reder (eds.). Chicago: Univ. of Chicago Press, pp. 67-94.

Can Trailers Help to Alleviate Popularity Bias in Choice-Based Preference Elicitation?

Mark P. Graus
Eindhoven University of Technology
IPO 0.20
PB 513, 5600 MB, Eindhoven, Netherlands
m.p.graus@tue.nl

Martijn C. Willemsen
Eindhoven University of Technology
IPO 0.17
PB 513, 5600 MB, Eindhoven, Netherlands
m.c.willemsen@tue.nl

ABSTRACT

Previous research showed that choice-based preference elicitation can be successfully used to reduce effort during user cold start, resulting in an improved user satisfaction with the recommender system. However, it has also been shown to result in highly popular recommendations. In the present study we investigate if trailers reduce this bias to popular recommendations by informing the user and enticing her to choose less popular movies. In a user study we show that users that watched trailers chose relatively less popular movies and how trailers affected the overall user experience with the recommender system.

CCS Concepts

•Information systems → Recommender systems; •Human-centered computing → Human computer interaction (HCI); User studies; *User models;*

Keywords

Recommender Systems; Choice-based Preference Elicitation; User Experience; Trailers; Information

1. INTRODUCTION

1.1 Cold Start

New user cold start is one of the central problems in recommender systems. It occurs when a user starts using a recommender system. As there is no information for this user to base recommendations on, the recommender system requires her to provide feedback in order to receive recommendations. This requires quite often significant effort of the user.

In addition, as users watch only a certain amount of movies over any time period, asking users to provide a set amount of feedback may require them to provide feedback on items that they have experienced a longer time ago which will require them to rely on memory. This can lead to unreliable feedback [2].

1.2 Choice-Based Preference Elicitation

One way to reduce the effort can be found in choice-based preference elicitation. Where most recommender systems ask users to provide a number of ratings on items (explicit feedback), recommender systems applying choice-based preference elicitation ask the user to make a number of choices (implicit feedback). Using implicit feedback to produce personalized ranking has been shown to provide better fitting prediction models than using explicit feedback [8]. In recent user studies users of collaborative filtering systems were provided with choice-based preference elicitation[4, 6]. Where in the more standard rating-based preference elicitation people are asked to rate the items they know, in choice-based preference elicitation they are asked to choose the item that best matches there preference from a list. In our own work, this alternative has been shown to require less effort than more standard rating-based preference elicitation, while allowing for more control, resulting in more novel and accurate recommendations [4].

Other work compared a recommender system using ratings against a recommender system using pair-wise comparisons (i.e. choices between two alternatives)[1]. The system using comparisons provided better recommendations in terms of objective performance metrics (nDCG and precision). In addition, users preferred the system using pair-wise comparisons as it made them more aware of their choice options and provided a nicer user experience.

One observation in [4] was that providing users with the possibility to indicate their preferences through choices resulted in a bias towards more popular movies, and subsequently users received more popular recommendations. Although this experiment showed that popularity leads to higher satisfaction at that moment in the lab setting of the study, such popular recommendations may not provide sufficient value in normal, long term usage scenarios.

1.3 Memory Effects in Recommender Systems

Memory effects could be a possible explanation for this bias towards popular movies that results in people receiving overly popular recommendations. In rating-based recommender systems memory effects have been shown to influence how users provide feedback. Bollen et al.[2] have demonstrated that ratings given closer to the time the movie was actually watched tend to be more extreme than ratings for movies that have been watched a longer time ago. They

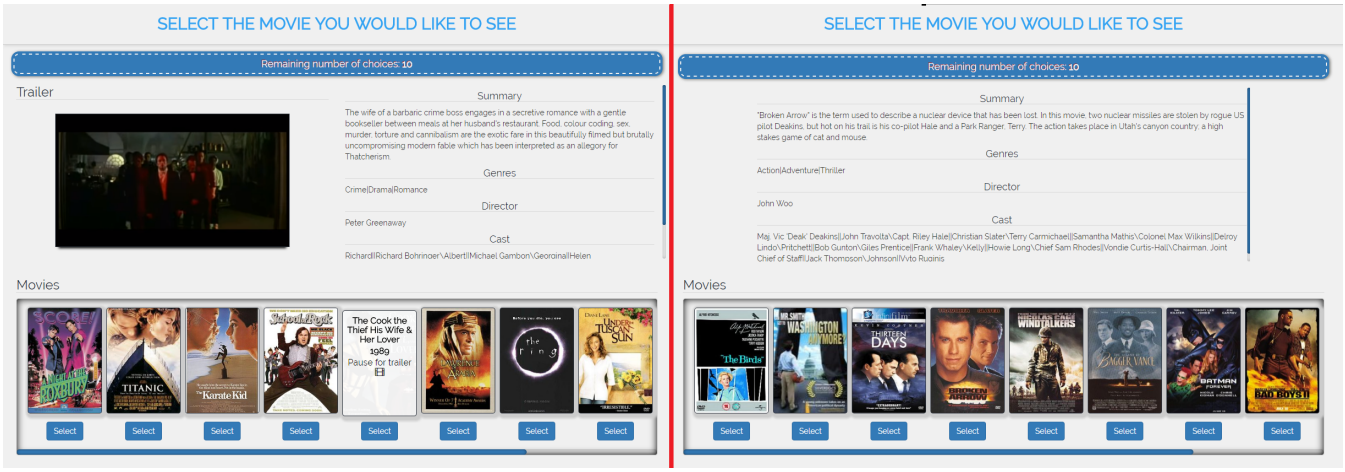


Figure 1: User Interface used for the study. The interface to the left is used for participants in the trailer condition, the interface to the right for participants in the non-trailer condition. Within the interface the list of items to choose from is shown below, the trailer and additional metadata is above.

argue that this is because of people forgetting information about the movies required to rate them, which has consequences for the reliability of the input provided. This same effect could result in users choosing items that they recognize in a choice-based preference elicitation task: it is more likely that people remember more popular movies than less popular movies.

1.4 Trailers as source of extra information

The current study tries to investigate if this bias towards picking popular movies can be alleviated by giving users additional information to make more informed choices.

In order to both minimize the effort required and maximize the reliability of the input given during the new user cold start situation, we propose to use choice-based preference elicitation and provide the user with additional information to give her the means to make more informed choices.

In most recommender systems users can already rely on meta-information like for example genre, cast and a synopsis. A possible additional source of information about a movie can be found in trailers. Trailers may help the user in two ways. Firstly, trailers can help a user in refreshing the memory to provide reliable feedback, alleviating potential memory problems described in the previous section. Secondly, even for movies that a user has not seen yet, a trailer can be used to evaluate whether or not a movie is worth watching. This is an advantage of choice-based preference elicitation over rating-based preference elicitation, because in rating-based users typically only rate (and provide information on) movies they have actually watched.

1.5 Research Question and hypotheses

The present research aims to investigate how providing additional information in the form of trailers during choice-based preference elicitation affects the interaction in terms of both objective behavior and subjective user experience.

In terms of objective behavior we hypothesize that trailers allow users to make more informed choices and rely less on popularity when making these choices. In other words, we expect the possibility to watch trailers to reduce the popularity of the items a user chooses.

In terms of user experience we expect trailers to provide the user with more information, which is expected to be reflected in the perceived informativeness of the system. As we expect trailers to motivate users to select less popular movies, we expect perceived recommendation novelty (the opposite of popularity) and diversity to increase. Both novelty and diversity may affect system and choice satisfaction.

It is hard to formulate expectations about the direction of the effect of trailers on user satisfaction. We expect user satisfaction in this setting to consist of system satisfaction (i.e. "how well does this system help me") and choice satisfaction ("how happy am I with the item that I choose based on this system"). In previous research novelty and system satisfaction were shown to be negatively correlated [9, 4]. On the other hand, trailers might make users open to less popular movies and as such novelty could have a positive effect on choice satisfaction. Additionally previous studies have shown that system satisfaction positively influences choice satisfaction[9]. Having the possibility to watch trailers may result in an increased system satisfaction and thus choice satisfaction. Considering all these effects it is hard to foresee in what way trailers will affect user experience.

The expected effects are shown in Fig. 2 below, with where possible the directions of the hypothesized effect.

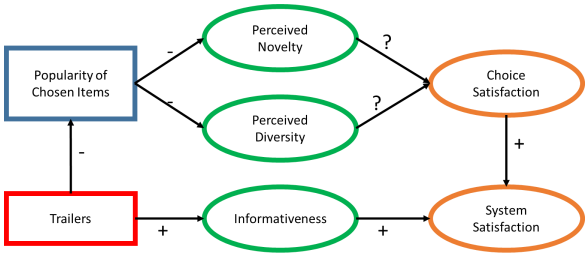


Figure 2: Path diagram of expected effects.

Table 1: Texts used for the items in the survey, with item factor loadings and factor robustness per aspect of user experience.

Considered Aspect	Item	Factor Loading
Informativeness AVE = 0.587 $\alpha = 0.71$	I got sufficient information on each movie to make a choice.	
	Visual information is more important to me for making a choice than written information.	
	I like the way information about the movies is provided to me in this system.	0.936
	The system provided too much information for each movie.	
	I would rather have different information about the movies than what I got from the system to make a choice	-0.647
Diversity AVE: 0.655 $\alpha = 0.80$	The recommendations contained a lot of variety.	
	The recommendations covered many movie genres.	
	All the recommended movies were similar to each other.	0.822
	Most movies were from the same genre.	0.867
	The recommended list of movies suits a broad set of tastes.	-0.798
Novelty	The recommended list of movies has a lot of movies I did not expect.	
	The recommended list of movies has a lot of movies that are familiar to me.	
	The recommended list of movies has a lot of pleasantly surprising movies.	
	The recommended list of movies has a lot of movies I would not have thought to consider.	
	The recommender provides few new suggestions.	
System Satisfacation AVE: 0.814 $\alpha = 0.88$	I like using the system.	0.913
	Using the system is a pleasant experience.	0.935
	I would recommend the system to others.	0.859
	The system is useless.	
	The system makes me more aware of my choice options.	
	I can find interesting items using the recommender system.	
Choice Satisfacation AVE: 0.692 $\alpha = 0.81$	I like the movie I've chosen from the final recommendation list.	0.820
	I was excited about my chosen movie.	
	The chosen movie fits my preference.	0.753
	I know several items that are better than the one I selected.	
	My chosen movie could become part of my favourites.	
	I would recommend my chosen movie to others/friends.	0.932

2. METHOD

A system was developed to address the research questions through an online study. Participants were invited to browse to a website where they could access our recommender system. Upon entering the website participants were assigned randomly to one of two experimental conditions: the trailer condition, where participants were given the possibility to watch trailers and the non-trailer condition, where participants could not watch those trailers. They were subsequently shown an introduction page with an informed consent form and a brief explanation about the task at hand.

After the explanation, the preference elicitation phase started (see Fig 1 for a screenshot), where the experimental manipulation came into effect. Participants in the trailer condition were able to see trailers, where participants in the non-trailer condition were not. Applying the same methodology as in [4] participants were presented with a set of 10 movies to choose from. The participants in the trailer condition would be informed about how they could watch trailers for the recommended movies. Participants were asked to evaluate the list and select the movie they would like to watch.

After choosing, the system would incorporate the choice and provide the participant with a new set of recommendations. Participants would be assigned a null vector upon entering. After which each choice was incorporated by the recommender system in four steps, described in more detail in [4]. Firstly, the user vector in the matrix factorization model was moved in the direction of the chosen item. Secondly, new rating predictions were calculated. Thirdly, the proportion of movies with lowest predicted rating was discarded. Fourthly a new choice set was calculated by taking the maximally diversified set from the remaining movies. Diversification was done through a greedy selection algorithm [7] with the goal of minimizing intra-list similarity [3] by maximizing the sum of the distances in the matrix factorization space between recommended items.

After 9 such choices, the user would see an explanation about how the choices they made would be used to calculate the final set of recommendations. The screen with final recommendations was identical to the previous screens except for the explanation. The final recommendations consisted of the Top-10 movies based on the last calculated user vector. People were asked to make the final choice from this list after which they were invited to complete a survey designed to measure the user experience.

The interface allowed users to watch trailers in the trailer condition by hovering over the presented movie covers. The trailers were retrieved through The Movie Database[1]. After hovering for 2 seconds, a video player would appear in allocated space in the interface. Each trailer for which a user

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

pressed the play button was stored as a view.

2.1 Recommender Algorithm

The recommendations were predicted through a matrix factorization model trained on ratings for the 2500 most rated movies in the 10M MovieLens dataset. The final dataset consisted of 69k users, 2500 items and 8.82M ratings. The performance metrics of the used model were up to standards (MAE: 0.61358, RMSE: 0.79643, measured through 5-fold cross-validation).

2.2 Participants

In total 89 participants made at least one choice in the system. Participants were recruited from different courses in the department and were entered in a raffle for one of 5 gift cards. No demographic information was asked. Out of the 89 participants 50 were in the condition where no trailers could be watched, 39 were able to watch trailers. The people who were able to do so, watched on average 10.38 trailers ($median = 10, SD = 9.69$).

In total 74 participants completed the survey. After inspection, data from 3 participants was removed because they completed the survey unrealistically fast. A total of 71 (40 of which did not have the possibility to watch trailers, 31 did) responses was thus used to study the effects on user experience.

2.3 Measures

In order to test our hypothesis we measured aspects of behavior and developed a survey to measure user experience. In terms of behavior we measured the popularity of the movies people chose and whether or not they watched any trailers. Popularity is defined as the ranked order based on the number of ratings in the original MovieLens dataset. The movies are ranked from the most rated (1) to the least rated (2500).

We investigate the user experience following the evaluation framework from Knijnenburg et al. [5]. In line with the research questions we developed a survey with the aim of measuring 5 aspects of user experience: perceived informativeness, perceived novelty, perceived diversity, system satisfaction and choice satisfaction. The items used are shown in Table 1. All items were submitted to a confirmatory factor analysis (CFA). The CFA used repeated ordinal dependent variables and a weighted least squares estimator, estimating 5 factors. Items with low factor loadings, high cross-loadings, or high residual correlations were removed from the analysis. The factor loadings for the novelty construct were not sufficiently high, so it was dropped from the final factor analysis.

3. RESULTS

The results section will first describe how trailers affect the choices users make. After that the analysis of the survey data will provide insight in how trailers affect the user experience.

3.1 Behavior

The effects on user behavior are expected to be two-fold. Firstly, as trailers allow the user to make more informed choices, we expect the individual chosen items to be less popular for people watching trailers. In other words, movies chosen by users who watch trailers are expected to have lower popularity ranks. Secondly, when people make less popular choices throughout the interaction with the system, we expect the individual choice sets to be less popular as a whole. For users that watch trailers we expect the average popularity rank of choice sets is expected to be lower.

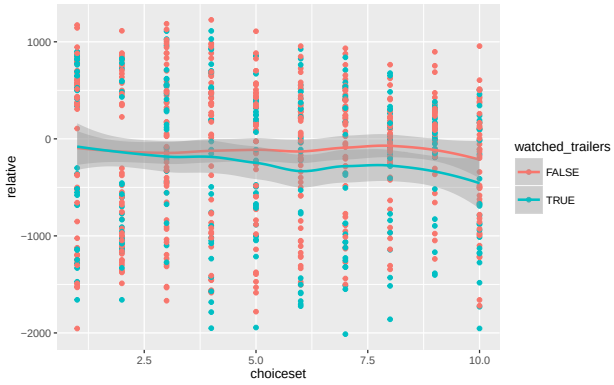


Figure 3: Relative Rank of Chose Item as a Function of Choice Set, for participants that watched trailers (Green) and those that did not (Red).

An alternative way to study this effect is by looking at the relative popularity of the choices users make, instead of the absolute popularity. To do this we calculated for each choice the difference between the popularity rank of the chosen item and the average popularity rank of the items in the set. If this number is positive, the chosen item is above average in terms of popularity, if it is negative, the chosen item is below average in terms of popularity.

Although there was no difference across experimental conditions, the plot in Figure 3 shows that for participants that actually watched trailers (i.e. people in the trailer conditions that watched at least one trailer) the relative popularity of the chosen item decreases after around 5 choices compared to participants that did not watch trailers (i.e. people in the non-trailer condition or in the trailer condition that did not watch any trailers). In a repeated measures ANOVA this effect proves to be significantly lower ($F(1, 87) = 6.992, p < 0.01$) for users that watched trailers. Watching trailers thus made users choose relatively less popular movies.

In order to understand what the results are for the user experience we investigate the survey data.

3.2 User Experience

The subjective constructs from the CFA were organized into a path model using Structural Equation Modeling. The resulting model had good model fit ($\chi^2(66) = 1052.974, p < 0.001, CFI = .997, TLI = .996, RMSEA = .029, 90\%CI : [0.000, 0.084]$). The corresponding path model is shown in Figure 4.

Different from earlier studies[5] we did not find that system satisfaction influences choice satisfaction directly. Moreover, system and choice satisfaction are not strongly related. A possible explanation for this could be that in this study the distinction between the preference elicitation task and recommendation stages is less clear than in previous studies. As every choice task has the same appearance as a set of recommendations (despite the clear explanation), the choice

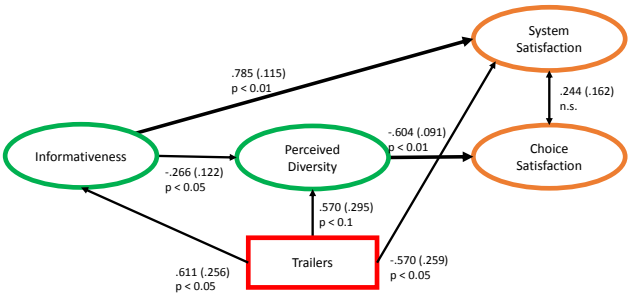


Figure 4: Path model of the CFA. Width of the arrows show effect sizes, numbers next to the arrows show the standardized effect size, with standard error and significance levels.

Figure 5: Boxplots of the estimated marginal means for the perceived informativeness (INF), perceived diversity (DIV), system satisfaction (SYSSAT) and choice satisfaction (CHOSAT), for participants in the trailer (red bars) and non-trailer (green) conditions.

task from the final list of recommendations might not have been perceived as much different from the choice tasks during the preference elicitation task. System Satisfaction in turn is positively influenced by Informativeness. In addition, the more people experience Informativeness, the less they perceive Diversity. Opposed to previous studies, we find that higher diversity results in a lower Choice Satisfaction.

In order to investigate the overall effects of the trailers we in addition consider the marginal means. The trailers affect the user experience in a number of ways. Firstly, providing trailers is experienced as an increase in informativeness of the system (statistically significant: $\beta = 0.664, t(69) = 3.142, p < 0.01$), as can be seen in the path model (Figure 4) and the marginal means (Figure 5).

It also results in an increased perceived diversity, but this effect is counteracted by the decrease as a result of the increased informativeness. This indirect effect of the manipulation on diversity through perceived informativeness results in the non-significant effect we observer in Figure 5. As far as system satisfaction is concerned, trailers actually decrease the system satisfaction. But similar to perceived diversity, this direct effect of trailers on system satisfaction is counteracted by the positive effect of the increased informativeness on system satisfaction.

4. CONCLUSION AND DISCUSSION

The present study aimed to decrease the tendency to use popularity as a heuristic in a choice-based preference elicitation task by providing users with means to make informed choices.

The analysis on user behavior showed that people watching trailers are more inclined to pick relatively less popular items. By investigating the user experience we found that aside from the impact on the decisions users make, the user experience was influenced. Informativeness of the system increased with the possibility of watching trailers. While no significant differences were found on the other aspects of user experience, the path model provides insight in the positive and negative consequences of providing trailers, consisting of an increased informativeness and diversity, but decreased system satisfaction.

4.1 Limitations

One of the limitations is that the effect of trailers on user experience with a recommender system is not tested against a more standard approach of preference elicitation. As users expressed rating items costs more effort than choosing[4], providing them with trailers during rating tasks may make the task cost too much effort and subsequently users may decide to not look at trailers. Nonetheless, comparing the effects of using trailers in choice-based versus rating-based preference elicitation can be valuable future research.

One aspect of behavior worth investigating based on the findings in this study is information regarding in what stage of the preference elicitation task users watched trailers. The way data was stored in the current dataset does not allow us to investigate for example if people watch more trailers in the beginning, or towards the end of the study, which could provide more fine grained insight in how trailers influence the choices people make. Future research should incorporate not only whether or not people watch trailers, but also when they do so. It would be particularly interesting to see if users use trailers differently in the choice of the final recommendations compared to the choices during the preference elicitation task.

The effect of popularity on choice satisfaction needs to be investigated in more detail. Previous studies have shown that popularity of recommendations has a positive influence on choice satisfaction in lab settings, but whether or not this effect remains in the long run needs to be investigated. It is possible that popularity can be used as a heuristic when evaluating a recommender system, but that longer term interaction is actually harmed by high popularity.

Acknowledgments

We thank Olivier van Duuren, Niek van Sleeuwen, Bianca Ligt, Daniëlle Niestadt, Shivam Rastogi and Suraj Iyer for programming the user inferface and performing the experiment as part of their student research project.

5. REFERENCES

[1] L. Blédaité and F. Ricci. Pairwise preferences elicitation and exploitation for conversational collaborative filtering. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 231–236, New York, NY, USA, 2015. ACM.

[2] D. Bollen, M. Graus, and M. C. Willemsen. Remembering the stars? In *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*, page 217, New York, New York, USA, 2012. ACM Press.

[3] P. Castells, N. J. Hurley, and S. Vargas. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, volume 54, pages 881–918. Springer US, Boston, MA, 2015.

[4] M. P. Graus and M. C. Willemsen. Improving the User Experience during Cold Start through Choice-Based Preference Elicitation. In *Proceedings of the 9th ACM Conference on Recommender Systems - RecSys '15*, pages 273–276, New York, New York, USA, 2015. ACM Press.

[5] B. Knijnenburg, M. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[6] B. Loepp, T. Hussein, and J. Ziegler. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 3085–3094, 2014.

[7] E. Minack, W. Siberski, and W. Nejdl. Incremental diversification for very large sets. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 585, 2011.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, United States, 2009. AUAI Press.

[9] M. C. Willemsen, M. P. Graus, and B. P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *UMUAI*, under revision.

Scalable Exploration of Relevance Prospects to Support Decision Making

Katrien Verbert
Department of Computer Science
KU Leuven
Leuven, Belgium
katrien.verbert@cs.kuleuven.be

Karsten Seipp
Department of Computer Science
KU Leuven
Leuven, Belgium
karsten.seipp@cs.kuleuven.be

Chen He
Department of Computer Science
KU Leuven
Leuven, Belgium
chen.he@cs.kuleuven.be

Denis Parra
Dept. of Computer Science
Pontificia Universidad Católica de Chile
Santiago, Chile
dparras@uc.cl

Chirayu Wongchokprasitti
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA, USA
chw20@pitt.edu

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT

Recent efforts in recommender systems research focus increasingly on human factors that affect acceptance of recommendations, such as user satisfaction, trust, transparency, and user control. In this paper, we present a scalable visualisation to interleave the output of several recommender engines with human-generated data, such as user bookmarks and tags. Such a visualisation enables users to explore which recommendations have been bookmarked by like-minded members of the community or marked with a specific relevant tag. Results of a preliminary user study (N=20) indicate that effectiveness and probability of item selection increase when users can explore relations between multiple recommendations and human feedback. In addition, perceived effectiveness and actual effectiveness of the recommendations as well as user trust into the recommendations are higher than a traditional list representation of recommendations.

CCS Concepts

•**Human-centered computing** → **Information visualisation; Empirical studies in visualisation;** *User interface design;*

Keywords

Interactive visualisation; recommender systems; set visualisation; scalability; user study

1. INTRODUCTION

When recommendations fail, a user's trust in a recommender system often decreases, particularly when the system acts as a "black box" [7]. One approach to deal with this issue is to support exploration of recommendations by exposing recommendation mechanisms and explaining why a certain item was selected [19]. For example, graph-based visualisations can explain collaborative filtering results by representing relationships among items and users [11, 3].

Our work has been motivated by the presence of *multiple relevance prospects* in modern social tagging systems. Items bookmarked by a specific user offer a *social relevance prospect*: if this user is known or appears to be like-minded, a collection of her bookmarks is perceived as an interesting set that is worth to explore. Similarly, items marked by a specific tag offer a *content relevance prospect*. In a social tagging system extended with a personalised recommender engine [12, 15, 4], top items recommended to a user offer a *personalised relevance prospect*.

Existing personalised social systems do not allow their users to explore and combine these different relevance prospects. Only one prospect can be explored at any given time: a list of items suggested by a recommender engine, a list of items bookmarked by a user, or a list of items marked with a specific tag. In our work, we focus on the use of visualisation techniques to support exploration of *multiple* relevance prospects, such as relationships between different recommendation methods, socially connected users, and tags, as a basis to increase acceptance of recommendations. In earlier work, we investigated how users explore these recommendations using a cluster map visualisation [20]. Although we were able to show the potential value of combining recommendations with tags and bookmarks of users, the user interface was found to be challenging. Further, the nature of the employed visualisation made our approach difficult to scale: in a field study, users only explored relations between a maximum of three entities. Due to these limitations, the effect of using multiple prospects could not be fully assessed.

In this paper, we present the use of a *scalable* visualisation that combines personalised recommendations with two additional prospects: (1) bookmarks of other users (a *social* relevance prospect), and (2) tags (*content* relevance prospect). Personalised recommendations are generated with four different recommendation techniques and embodied as *agents*

to put them on the same ground as users (i.e., recommendations made by agents are treated in the same way as bookmarks left by users). We use the UpSet visualisation [9], which offers a scalable approach to combine multiple sets of relevance prospects, i.e. different recommender agents, bookmarks of users, and tags. We aim to assess whether the combination of multiple relevance prospects shown with this technique can be used to increase the effectiveness of recommendations while also addressing several issues related to the "black box" problem. In particular, we explore the following research questions:

- **RQ1** Under which condition may a scalable visualisation increase *user acceptance* of recommended items?
- **RQ2** Does a scalable set visualisation increase *perceived effectiveness* of recommendations?
- **RQ3** Does a scalable set visualisation increase *user trust* in recommendations?
- **RQ4** Does a scalable set visualisation improve *user satisfaction* with a recommender system?

The contribution of this research is threefold:

1. First, we present a novel interface that integrates a simplified version of the UpSet visualisation, allowing the user to flexibly combine multiple prospects to explore recommended items.
2. Second, we present a preliminary user study that assesses the effect of combining multiple relevance prospects on the decision-making process. We find that users explore combinations of recommendations with users and tags more frequently than recommendations only based on agents. Further, this combination is found to provide more relevant results, leading to an increase in user acceptance.
3. Third, we find indications of an increase in user trust, user satisfaction, and both perceived and actual effectiveness of recommendations compared to a baseline system. This shows the positive effects of combining multiple prospects on user experience.

This paper is organized as follows: first, we present related work in the area of interactive recommender systems. We then introduce the design of IntersectionExplorer, an interactive visualisation that allows users to explore recommendations by combining multiple relevance prospects in a scalable way. We assess its impact on the decision-making process and finish with a discussion of the results.

2. RELATED WORK

In a recent study, we analyzed 24 interactive recommender systems that use a visualisation technique to support user interaction [6]. A large share of these systems focuses on transparency of the recommendation process to address the "black box" issue. Here, the overall objective is to explain the inner logic of a recommender system to the user in order to increase acceptance of recommendations. Good examples of this approach are PeerChooser [11] and SmallWorlds [5]. Both allow exploration of relationships between recommended items and friends with a similar profile using multiple aspects.

In addition, TasteWeights [3] allows users to control the impact of the profiles and behaviours of friends and peers on the recommendation results. Similar to our work, TasteWeights provides an interface for such hybrid recommendations. The system elicits preference data and relevance feedback from users at run-time in order to adapt recommendations. This idea can be traced back to the work of Schafer et al. [17] concerning meta-recommendation systems. These meta-recommenders provide users with personalised control over the generation of recommendations by allowing them to alter the importance of specific factors on a scale from 1 (not important) to 5 (must have). SetFusion [13] is a recent example that allows users to fine-tune weights of a hybrid recommender system. SetFusion uses a Venn diagram to visualise relationships between recommendations. Our work extends this concept by visualising relationships between different relevance prospects, including human-generated data, such as user bookmarks and tags in addition to outputs of recommenders, in order to incite the exploration of related items and to increase their relevance and importance in the eye of the user. To do so, we employ a set-based visualisation that allows users to quickly discern relations and commonalities between the items of recommenders, users, and tags for a richer and more relevant choice.

Relevance-based or set-based visualisation attempts to spatially organize recommendation results. This type of visualisation has its roots in the field of information retrieval and was used for the display of search results. For example: for a query that uses three terms, this type of visualisation would create seven set areas. Three sets will show the results separately for each term. Another set of three will show results for any combination of two of these terms. Finally, one set will show results that are relevant to all three terms together. The classic example of such a set-based relevance visualisation is InfoCrystal [18]. The Aduna clustermap visualisation [1] also belongs to this category, but offers a more complex visualisation paradigm and a higher degree of interactivity. The strongest point of both approaches, however, is the clear representation of the query terms and their relevant items, separately or in combination.

In the context of similar work, the novelty of the approach suggested in this paper is twofold: first, we use a set-based relevance approach that is not limited to keywords or tags, but which combines these with other relevance-bearing entities (users and recommendation agents). The major difference and innovation of our work is that we allow end-users to combine *multiple* relevance prospects to increase richness and relevance of recommendations. Second, we present and evaluate the use of a novel scalable visualisation technique (UpSet [9]) to perform this task and thereby demonstrate this approach's ability to increase recommendation effectiveness and user trust.

3. INTERSECTIONEXPLORER

IntersectionExplorer (IE) is an interactive visualisation tool that enables users to combine suggestions of recommender agents with user bookmarks and tags in order to find relevant items. We describe the visualisation and interaction design of the system, followed by its implementation.

3.1 Set Visualisation Design

We have adapted the UpSet [9] technique to visualise relations between users, tags, and recommendations. UpSet

represents set relations in a matrix view: while columns represent sets of different entities (such as recommender agents or other users' bookmarks), rows represent commonalities between these (Figure 1). The column header shows the name of the agent, user, or tag. The vertical bar chart below the column headers depicts the number of items belonging to each related set. Set relations are represented by the rows. In such a row, a filled cell indicates that the corresponding set contributes to the relation. An empty cell indicates that the corresponding set is not part of the relation. The horizontal bar chart next to each row shows the number of items that could be explored for this relation set. For example, the first row in Figure 1 indicates that there are three items that belong to both the set of recommendations suggested by the bookmark-based recommender agent, and the set of recommendations suggested by the tag-based agent. The second row shows suggestions of the bookmark-based agent only, whereas the third row only shows suggestions of the tag-based agent. For the convenience of the reader, we also depicted this relation in a traditional Venn diagram to support the understanding of the concept.

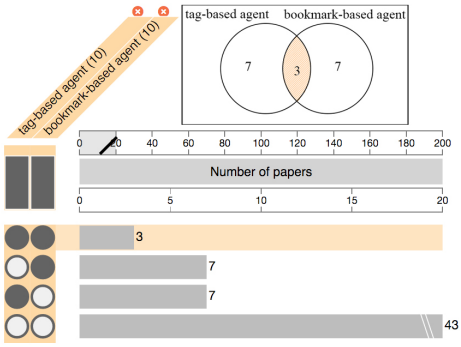


Figure 1: Set visualisation of IntersectionExplorer

One of the biggest advantages of a visual matrix is scalability. Whereas a Venn diagram can only display the intersections of a limited number of sets, the UpSet technique can present many sets in parallel, as only a single column has to be added to add another set to the visualisation. This greatly reduces space requirements while increasing the information density. The visual encoding of IE is identical for any number and constellation of sets. In practice, users may wish to first familiarise themselves with the display of a small number of sets, but due to the consistent and space-efficient design, they can seamlessly increase the set numbers without altering the view.

3.2 Interaction Design

An overview of the full IE interface is shown in Figure 2. The interface is separated into three connected parts. In the left part, the user can select different entities: agents, users and tags. If an agent is selected, the set of items suggested by this agent is added to the matrix visualisation in the canvas area. If a user is added, the set of bookmarks of this user is added. Similarly, if a tag is added, the set of papers marked with this tag is added to the view.

The canvas area represents user-selected sets as columns in a matrix view, allowing the user to explore overlaps between these sets. Each row represents relations between the different columns as explained in the previous section.

The user can explore the details of data items related to a certain row by clicking on the row. For example, after clicking the first row in Figure 2, the right part shows the title and authors of two papers that are bookmarked by "P Brusilovsky" and also suggested by three different agents.

The user can explore the items related to a specific set by clicking on the column header: all containing items of this set are then presented in the panel on the right. Meanwhile, the rows related to this set are also gathered at the top to facilitate exploration of relations with other sets.

At the top of the set view, the user can also sort the rows (set intersections) by *number of items* or *number of related sets* in ascending or descending order. The example of Figure 2 sorts the rows by the number of related sets in descending order. The first row represents items in the intersection of four sets. The second row represents items in the intersection of three sets and the next five rows represent items in the intersection of two sets. The other rows represent items related to a single set only.

3.3 Implementation

We have implemented IE on top of data from Conference Navigator 3 (CN3). CN3 is a social personalised system that supports attendees at academic conferences [14]. The main feature is its conference scheduling system where users can add talks of the conference to create a personal schedule. Social information collected by CN3 is extensively used to help users find interesting papers. For example, CN3 lists the most popular papers, the most active people, and the most popular tags assigned to the talks. When visiting the *talk page*, users can also see who scheduled each talk during the conference and which tags were assigned to this talk.

We use the list of conference talks as data items in IE. CN3 offers four different recommendation services that rely on different recommendation engines. The *tag-based* recommender engine matches user tags (tags provided by the user) with item tags (tags assigned to different talks by the community of users) using the Okapi BM25 algorithm [10]. The *bookmark-based* recommendation engine builds the user interest profile as a vector of terms with weights based on TF-IDF [2] using the content of the papers that the user has scheduled. It then recommends papers that match this profile of interests. Another two recommender engines, *external bookmark* and *bibliography*, are the augmented version of the *bookmark-based* engines [21]. The *external bookmark* recommender engine combines both the content of the scheduled papers and the research papers bookmarked by the user in academic social bookmarking systems such as Mendeley, CiteULike, or BibSonomy. Similarly, the *bibliography* recommender engine uses the content of papers published by the user in the past to augment the bookmarked papers.

The suggestions of these four recommender engines are represented as separate *agents* in IE. Users can explore which items are suggested by a single agent, for instance the tag-based recommender, but they can also explore which items are recommended by multiple agents to filter out the potentially more relevant recommendations. In addition, users can explore relations between agent suggestions and bookmarks of real users. As shown in Figure 2, the third row represents items suggested by the tag-based agent that have also been bookmarked by "P Brusilovsky", but that are not suggested by the two other agents and that have not been bookmarked by the active user ("K Verbert"). In this paper,

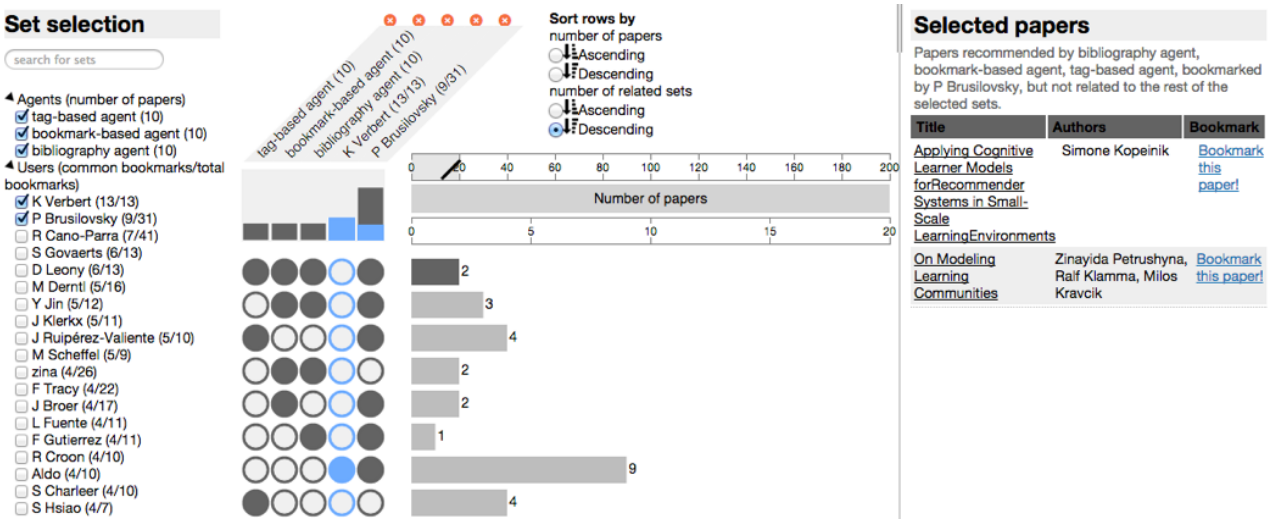


Figure 2: IntersectionExplorer visualises relationships between recommendations generated by multiple recommendation techniques (agents) and bookmarks of users and tags to increase relevance of recommendations.

we evaluate whether enabling users to explore relations between recommendations of different techniques, real users, and tags increases the acceptance of recommendations.

The set visualisation shows the relations of the selected sets as described in section 3.1. The column of the current user is displayed in blue while the other columns are represented in grey. As presented in Figure 2, the bar chart below the column headers of users overlays a blue bar that encodes the number of common bookmarks with the current user. The similarity between users is also represented next to the user name in the panel on the left: "P Brusilovsky (9/31)" indicates that the user "P Brusilovsky" has 31 bookmarks in total. Nine out of these 31 talks are also bookmarked by the active user ("K Verbert").

For the user study presented in this paper, we used the data from the EC-TEL conferences of 2014 and 2015. EC-TEL is a large conference on technology enhanced learning. We retrieved user bookmarks and tags of these conferences, and had access to the different recommender services for both the 2014 and 2015 edition of the conference. Attendees of the EC-TEL conference participated in the user study that is presented in the next section.

4. USER STUDY

To investigate to what extent the set visualisation may support users in finding relevant items, we conducted a within-subjects study with 20 users (mean age: 32.9 years; SD: 6.32; female: 3) in two conditions, both of which had to be completed by all participants.

In the first condition (baseline), users were tasked to explore recommendations presented to them using the CN3 "my recommendations" page with four ranked lists. In the second condition, users explored recommendations using IntersectionExplorer (IE). To avoid a learning effect, each condition used a separate data set from which to generate recommendations. The baseline condition (CN3) used the EC-TEL 2014 proceedings (172 items), the IE condition used the EC-TEL 2015 proceedings (112 items).

To prepare for the study, users bookmarked and tagged five items in each of the proceedings. In addition, users' publication history and academic social bookmark systems (CiteULike and Bibsonomy) were read. From the combined data, recommendations were generated in both conditions using the four different techniques described in Section 3. These were then presented as four individual agents: a tag-based agent, a bookmark-based agent, an external bookmark agent and a biography agent.

To explore the impact of the IE visualisation on the users' acceptance of items, users were tasked to explore the recommendations of the four agents freely and to bookmark five items. During this period we recorded the time and amount of steps taken to create a bookmark. In particular, we recorded the following actions: selection/deselection of agents, users and tags, sorting, hovering over a result row (if mouse position was held for more than two seconds), clicking onto a paper's title, and clicking the bookmark button. Further, we collected data using a think-aloud protocol, synchronizing screen recording and microphone input. Finally, users completed a questionnaire using a five-point Likert scale. The questions were based on ResQue [16] and the framework proposed by Knijnenburg et al. [8], both of which have been validated for the measurement of subjective aspects of user experience with recommender systems.

Before exploring the recommendations using IE, users were shown a three-minute video to explain the system's operation. In the IE view, users saw the intersections with the agents' recommendations. In the CN3 (baseline) view, users saw the full results of the bookmark agent and could navigate to the recommendations generated by the three other agents, as presented in Figure 3. The study was counterbalanced by mode of exploration (CN3/IntersectionExplorer). Five users completed the study with a researcher present in the same room, whereas 15 users completed the study via an on-line video call. To establish users' background-knowledge, we asked each participant a set of questions using a five-point Likert scale after the study. Mean results were as follows:

- Users were familiar with technology-enhanced learning

(mean: 4; SD: 1.1).

- Users were familiar with recommender systems (mean: 4; SD: 0.95).
- Users were familiar with visualisation techniques (mean: 4.05; SD: 0.86).
- Users occasionally followed the advice of recommender systems (mean: 4.25; SD: 0.77).
- Eight participants had never heard of CN3 before. Twelve had heard of it, but had no particular familiarity with the system (mean: 3.25; SD: 1.13).

One user had no publications, four had two to four publications, fifteen had five publications or more. Within the last group, 93.3% had published on an EC-TEL conference in the past.

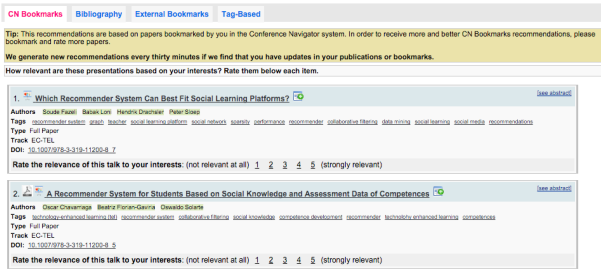


Figure 3: CN3 baseline interface with four ranked lists provided by four recommender engines.

4.1 Results and Evaluation

4.1.1 *Quantitative results*

The main focus of this study was to investigate under which condition the visualisation may increase user acceptance of recommended items. To answer our question, we need to analyse the in-depth behaviour of users exploring the recommendations using various combinations of recommender agents and the bookmarks and tags of other users.

In order to be able to determine the impact of visualising relations between agents, users, and tags, we defined two measures: effectiveness and yield.

Effectiveness measured how frequently the exploration of a specific set providing a number of intersections (henceforth called 'size') eventually led to the user bookmarking another paper (from the recommended set of papers). By the exploration of a set we mean clicking on a row of intersections in the visualisation (Figure 1, Figure 2) to show the items belonging to the intersection of the selected sets.

Effectiveness was calculated as the number of cases where the exploration of an intersection of a specific type and size resulted in a new user bookmark, divided by the number of times this intersection type and size was explored. Intersection types could be a single agent, a combination of agents, or a combination of agents with another entity (user or tag). The size represented the number of sets in the intersection. For instance, users explored suggestions of a single agent 26 times in total (one agent, Figure 4, first row). Exploration of these sets resulted in the creation of five bookmarks. Thus, the single agent's effectiveness is 5/26 = 19%.

Yield measured the fraction of items of an explored set that were actually bookmarked by all users in total. For instance, if the results of the intersection with one agent listed a total of 93 items for all users combined, but only five bookmarks were created from this type and size of intersection across the whole study, its overall yield was 5/93=0.05 (Figure 5, first row).

Figure 4 and Figure 5 reveal an interesting effect: sets which included the recommendations of agents and other entities, such as other users' bookmarks and tags, appeared to have a higher yield and effectiveness than sets based on agent recommendation alone, even if the number of intersections were the same. To further explore this aspect, we divided the results for effectiveness and yield into two groups: those obtained for interaction with one to four agents, and those obtained from interaction with the recommendations of different numbers of agents and another entity (user or tag). A Friedman test indicated a significant effect of recommendation source on effectiveness, $c^2(1) = 4$, $p = .046$ revealing that users who explored the recommendations of agents combined with another entity in the recommendation matrix of IE (median: .43), tended to find more than twice as many relevant items as when only using the agents for the recommendation (median: 0.21) (Figure 4). These results correspond to our findings that the richer the set (the more "perspectives" contribute the recommendation), the higher the yield (Figure 6). In general, Figure 7 shows that the larger the amount of intersections with a specific type, the higher the yield. Pearson's correlation showed a positive correlation between the number of intersections and yield (r = .839, n = 6, p = .037).

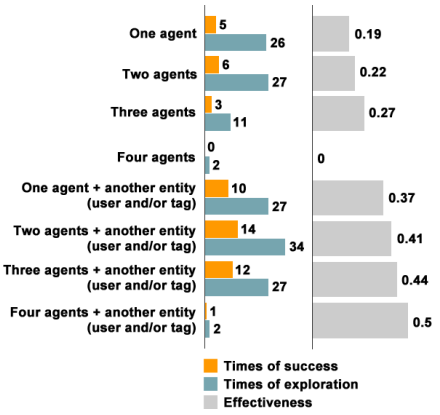


Figure 4: Effectiveness of the combinations of various amounts of agents and the combinations of various amounts of agents and other entities, such as users or tags. Effectiveness was higher when agents were combined with another entity.

Overall, these results suggests that enriching automated recommendations based on tags, previous bookmarks, publication history and academic social bookmarks with *socially collected* relevance evidence, such as the bookmarks made by other users of the same conference or a tag, greatly increases the relevance of recommendations, resulting in a higher acceptance rate.

Regarding the overall operability of IE, an ANOVA of task completion time showed an effect of task number $F(4, 44) = 20.5$, $p < .001$ on interaction time. However, a post-hoc

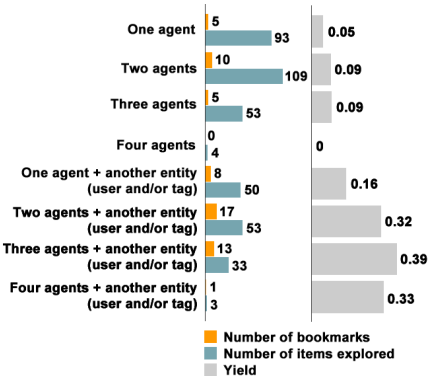


Figure 5: Yield of the combinations of various amounts of agents and the combinations of various amounts of agents and other entities, such as users or tags. Yield was higher when agents were combined with another entity.

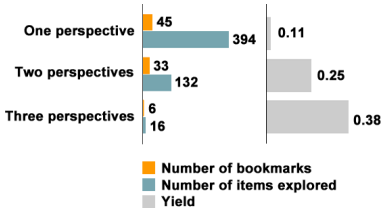


Figure 6: Yield of different numbers of perspectives in an exploration. Pearson's correlation showed a positive correlation between number of perspectives in an exploration and yield (r = 1.0, n = 3, p = .015).

Figure 7: Yield of different numbers of entities in the intersection. Pearson's correlation showed a positive correlation between number of entities in an intersection and yield (r = .839, n = 6, p = .037).

Bonferroni-Holm-corrected Wilcoxon signed-rank test indicated that differences were not statistically significant.

A Greenhouse-Geisser corrected ANOVA of the amount of steps needed to complete the bookmarking tasks showed an effect of condition, $F(1, 11) = 7.86$, $p = .017$, and an effect of task order, $F(2.09, 23) = 168.82$, $p = .002$. A Wilcoxon signed-rank test showed a trend for task one taking more steps when using IE (median: 11) than when using CN3 (median: 4), $Z = 2.5$, $p = .012$, but after applying a Bonferroni-Holm correction, differences were not statistically significant. This suggests that while IE may have a higher learning curve than CN3, no statistically significant differences exist in terms of efficiency of operation after acquaintance with the system (Figure 8).

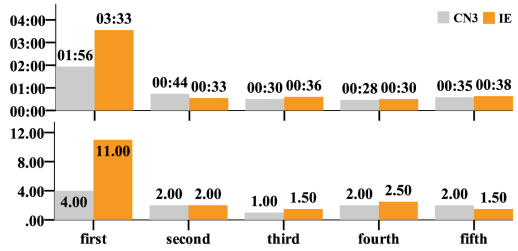


Figure 8: Median time (mm:ss) and steps of each task with IntersectionExplorer (IE) and CN3.

4.1.2 *Questionnaire results*

Results are reported in Figure 9, 10 and 11. Running a set of Bonferroni-Holm-corrected Wilcoxon signed-rank tests on the questionnaire results revealed the following:

- Papers explored with IE were perceived to be of a higher quality than with CN3 ($Z = 3.54$, $p < .001$).
- IE was perceived to be more effective than CN3 ($Z = 4.24$, $p < .001$).
- User satisfaction was higher with IE than with CN3 ($Z = 3.22$, $p = .001$).
- Users would me more willing to use IE frequently than CN3 ($Z = 3.42$, $p = .001$).
- Users perceived the recommendations shown in IE to be more trustworthy than CN3 ($Z = 2.55$, $p = .011$).

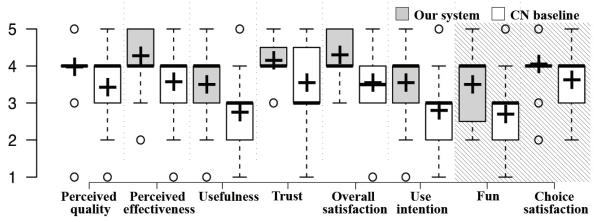


Figure 9: Questionnaire results with statistical significance. Differences between the aspects "Fun" and "Choice satisfaction" were not significant after the Bonferroni-Holm correction.

In addition, a trend was observed that users experienced IE to be more fun than CN3 ($Z = 2.28$, $p = .023$) and to provide a higher choice satisfaction ($Z = 2.1$, $p = .039$). However, after applying a Bonferroni-Holm correction, differences were not statistically significant.

Similarly, the results for the novelty of items (median: 4), effort to use the systems (median: 2), usefulness (median: 4), and ease of use (median: 4) were the same for both systems. Users tended to perceive the creation of bookmarks as more difficult in IE (median: 3) than in CN3 (median:

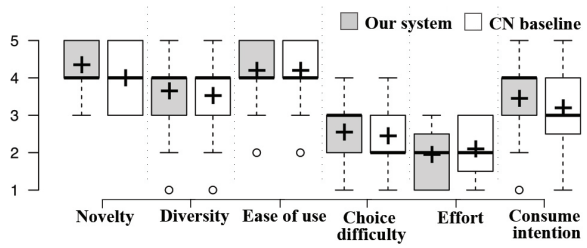


Figure 10: Questionnaire results without statistic significance.

2), but tended to read the bookmarked papers afterwards more frequently when using IE (median: 4) that when using CN3 (median: 3).

As for the IE-specific aspects shown in Figure 11, users perceived the visualisation to be adequate (median: 4) and the amount of information provided by the system to be sufficient to make a bookmark decision (median: 4). Users tended to be undecided regarding the interaction adequacy of IE (median: 3.5, see [16] for a definition), but found it easy to modify their preference to find relevant papers (user control, median: 4).

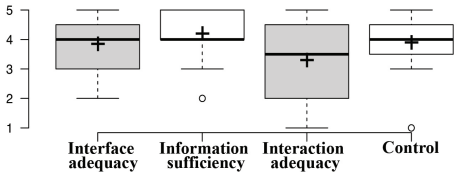


Figure 11: Interaction and visualisation sufficiency.

4.1.3 Observation

The think-aloud protocol revealed the following:

Interface: Some users misinterpreted empty circles to be a match of bookmarks or recommendations (three users) or initially failed to understand the meaning of the circles (three users). Others stated that they did not know that a tag-based agent was available and that the list of entities on the left was too long (three users).

Terminology: Two users had problems understanding the meaning of "sets", "related sets" or the numbers representing the amount of papers in a set.

It was further observed that some users only explored sets recommended by the agents, the majority explored sets recommended by agents and related to other users or tags.

4.1.4 Answering the research questions

RQ1 *Under which condition may a scalable visualisation increase user acceptance of recommended items?*

Our research showed that user acceptance of recommended items increased with the amount of sources used. However, the most important finding is that the addition of human-generated data – such as bookmarks of other users or tags – to the agent-generated recommendations resulted in a significant increase of effectiveness and yield. Our data suggests that providing users with insight into relations of recommendations with bookmarks and tags of community members increases user acceptance. We thus recommend to combine automated sources and personal sources whenever possible.

RQ2 *Does a scalable set visualisation increase perceived effectiveness of recommendations?*

Perceived effectiveness (expressed in the questionnaire) and actual effectiveness (how frequently users bookmarked a recommended paper) were increased by this type of visualisation.

RQ3 *Does a scalable set visualisation increase user trust in recommendations?*

The evaluation of the subjective data showed that user trust into the recommended items was increased with set-based visualisation of recommendation sources.

RQ4 *Does a scalable set visualisation improve user satisfaction with a recommender system?*

Overall, user satisfaction was higher when using the visualisation, suggesting this to be a key feature of the approach.

4.2 Discussion

4.2.1 Simplicity vs. effectiveness

The analysis of task completion time and amount of steps needed to complete the bookmarking tasks has shown that users require more time and interactions to set their first bookmark in IE, but that after this 'training phase', the operational efficiency between IE and CN3 does not differ. This corresponds to the observations made during the analysis of the think-aloud study, where it was found that some users initially struggled to understand the meaning of the different circle types or what a 'set' was.

However, the analysis of the subjective data has shown that users perceived IE to be more effective and its recommendations more trustworthy than those given by CN3. Especially the last point may be the result of removing the frequently lamented "black box" problem of recommenders by simply visualising how and why certain items are selected. In addition, users perceived items resulting from their use of IE to be of higher quality and found the overall experience more satisfying. This positive user experience may compensate for the initial conceptual problems encountered in the first exploration of the application and suggests that IE may be a helpful addition to the conference explorer service.

4.2.2 Comparison to previous work

In our previous work we presented the idea of combining recommendations embodied as agents with bookmarks of users and tags as a basis to increase effectiveness of recommendations [20]. A cluster map technique was used to enable users to explore these relations. Whereas the approach seemed promising, the cluster map was challenging for users to understand. In a first controlled user study, we asked users explicitly to explore recommendations of agents, bookmarks of users, tags and their combinations to try to find relevant items. Results of this user study indicate that there is an increase in effectiveness. In a follow-up uncontrolled field study users did not explore many intersections between different relevance prospects. As a result, the effect of combining relevance prospects could not be confirmed when users were not pushed to do so.

IE employs the novel UpSet visualisation technique that was presented at IEEE VIS in 2014. We simplified the interface and deployed it on top of data collected by Conference Navigator. The approach addresses the previous limitations

regarding ease of use and scalability: in this study users did explore many intersections, enabling us to investigate the effect of the approach on acceptance of recommendations.

4.2.3 Limitations

One limitation is the low number of participants. Further, the study was conducted with researchers from the field of technology enhanced learning with a high degree of visualisation expertise (mean: 4.05, SD: 0.86). Such users may be biased due to their graph literacy. In addition, our data was limited to that provided by the EC-TEL conferences.

5. CONCLUSION AND FUTURE WORK

We presented a study that used the UpSet visualisation technique to combine agent-based recommendations with human-generated recommendations in the form of bookmarks and tags. Despite the initial learning curve (when compared to the baseline system CN3), we found that this combination resulted in a higher degree of item exploration and acceptance of recommendations, than when using agent-only results. This way, user trust, usefulness, quality, and effectiveness were increased. We could thereby demonstrate the positive effects of the combination of multiple prospects on user experience and relevance of recommendations.

Future work will explore the applicability of our findings to a more diverse dataset and audience, as well as different types of visualisations. We have currently deployed IntersectionExplorer for attendees of the ACM IUI 2016 conference and will evaluate whether the visualisation can be used in an open setting, without the presence of a researcher. In addition, we plan to deploy the visualisation on top of data from large conferences, including the Digital Humanities conference series. Follow-up studies will assess the added value of our visualisation on top of larger data collections and with a less technical audience. With these studies, we intent to reach a wider range of users to further evaluate the effect of the approach on the effectiveness of recommendations.

6. ACKNOWLEDGMENTS

We thank all participants for their participation and useful comments. Part of this work has been supported by the Research Foundation Flanders (FWO), grant agreement no. G0C9515N, and the KU Leuven Research Council, grant agreement no. STG/14/019. The author Denis Parra was supported by CONICYT, project FONDECYT 11150783.

7. REFERENCES

[1] Aduna clustermap. www.aduna-software.com/technology/clustermap. Retrieved on-line 20 Augustus 2014.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[3] S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: a visual interactive hybrid recommender system. In *Proc. RecSys'12*, pages 35–42. ACM, 2012.

[4] J. Gemmell, T. Schimoler, B. Mobasher, and R. Burke. Recommendation by example in social annotation systems. In *E-Commerce and Web Technologies*, pages 209–220. Springer, 2011.

[5] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer. Smallworlds: visualizing social recommendations. In *Computer Graphics Forum*, volume 29, pages 833–842. Wiley Online Library, 2010.

[6] C. He, D. Parra, and K. Verbert. Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 2016.

[7] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proc. CSCW '00*, pages 241–250. ACM, 2000.

[8] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[9] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: visualization of intersecting sets. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1983–1992, 2014.

[10] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[11] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. Peerchooser: visual interactive recommendation. In *Proc CHI '08*, pages 1085–1088. ACM, 2008.

[12] D. Parra and P. Brusilovsky. Collaborative filtering for social tagging systems: an experiment with citeulike. In *Proc. RecSys '09*, pages 237–240. ACM, 2009.

[13] D. Parra and P. Brusilovsky. User-controllable personalization: A case study with setfusion. *International Journal of Human-Computer Studies*, 78:43–67, 2015.

[14] D. Parra, W. Jeng, P. Brusilovsky, C. López, and S. Sahebi. Conference navigator 3: An online social conference support system. In *UMAP Workshops*, pages 1–4, 2012.

[15] J. Peng, D. D. Zeng, H. Zhao, and F.-y. Wang. Collaborative filtering in social tagging systems based on joint item-tag recommendations. In *Proc CICM '10*, pages 809–818. ACM, 2010.

[16] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proc. RecSys'11*, pages 157–164. ACM, 2011.

[17] J. B. Schafer, J. A. Konstan, and J. Riedl. Meta-recommendation systems: User-controlled integration of diverse recommendations. In *Proc. CIKM '02*, pages 43–51, NY, USA, 2002. ACM.

[18] A. Spoerri. Infocrystal: A visual tool for information retrieval & management. In *Proc. CIKM '93*, pages 11–20. ACM, 1993.

[19] N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*, pages 479–510. Springer, 2011.

[20] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proc. IUI'13*, pages 351–362. ACM, 2013.

[21] C. Wongchokprasitti. *Using external sources to improve research talk recommendation in small communities*. PhD thesis, University of Pittsburgh, 2015.

Complements and Substitutes in Product Recommendations: The Differential Effects on Consumers' Willingness-to-pay

Mingyue Zhang
Tsinghua University
Beijing, China
zhangmy.12@sem.tsinghua.edu.cn

Jesse Bockstedt
Emory University
Atlanta, GA
bockstedt@emory.edu

ABSTRACT

Product recommendations have been shown to influence consumers' preferences and purchasing behavior. However, empirical evidence has yet to be found illustrating whether and how the recommendations of other products affect a consumers' economic behavior for the focal product. In many e-commerce websites, a product is presented with co-purchase and co-view recommendations which potentially contain complement and substitute products, respectively. Very little research has explored the differential effects of complementary and substitutable recommendations. In this study, we are interested in how the type of recommendations of other products impact the consumers' willingness-to-pay for the focal product, and additionally how the recommendations' price and the consumers' decision stage moderate this effect. We conducted a 2x2x2 randomized experiment to examine how the consumers' willingness-to-pay is affected by these factors. Experimental results provide evidence that there is no significant main effect difference between complementary and substitutable recommendations. But we observed a significant interaction effect between recommendation type and decision stage, which highlights the importance of timing in recommender systems. Other findings include that consumers are willing to pay more for a specific product when the price of a recommended product is high, as well as when they are in later decision stages. These findings have significant implications for the design and applications of recommender systems.

Keywords

recommender systems; complements; substitutes; willingness-to-pay; price; decision stage

1. INTRODUCTION

Recommender systems (RS) are becoming integral to how consumers discover new products and have a strong influence on what consumers buy and view. For instance, 60% of Netflix content consumption originates from recommendations, and 35% of Amazon sales are attributed to recommendations [11]. With their utmost importance for retailers, some studies have been conducted to explore the behavioral effects of recommender systems on consumers [1][2]. Specifically, prior studies found that consumers' preferences and willingness-to-pay for a product can be influenced by the values of personalized recommendations. This provides evidence that consumers' behavior is vulnerable toward recommendation agents. However, there is still space for researching the behavioral effects of detailed recommendation features. For example, when evaluating the information of a focal product on its own webpage, consumers are often exposed to additional relevant products as recommendations, such as on Amazon.com. It is an open research question whether consumer's purchase decisions such as willingness-to-pay for the focal product would be affected by the display of 'other products' and the type of information presented with these recommendations. Some work has studied this in offline settings [31], but little work addresses this issue in the online recommendation context.

Particularly, the types of 'other products' in a recommendation set may vary, but they can be generally categorized into substitutes and complements [19]. Substitutes are products that can be purchased instead of each other, while complements are products that experience joint demand. For example, when a user is evaluating a cellphone, it's reasonable to recommend other phones to better match his/her needs, but it also makes sense to recommend batteries, chargers, or cases, which commonly make up a bundle recommendation [34]. Research has shown that consumers factor into consideration the source or type of information when making their purchase decisions [26]. Economic theory suggests that complements increase demand for the focal product because of increasing the possibility of users finding added value for the focal product [31]. With the increased demand, the market price will increase accordingly, leading to higher individual willingness-to-pay. Whereas, substitutes decrease demand for the focal product due to competition, which leads to a lower market price and individual willingness-to-pay. Despite the extensive literature about complements and substitutes in economics, little research studied their differential effects in online recommendation settings.

In this study, we are interested in how the type of recommendations of other products impact the focal product, and additionally how the recommendations' price and consumers' decision stage moderate this effect. The price of a recommended product can be perceived as a contextual reference point. Along with the nature of cross-price elasticity of demands between complementary/ substitutable goods [20], the research question of how the price interacts with recommendation type is of great value. Additionally, when shopping online consumers tend to have a two-stage decision making process [9]: first, screening a large set of available products to identify a subset of the most promising alternatives; and second, evaluating the latter in more depth to make a final purchase decision. Zheng et al. (2009) [33] argue that consumers prefer different recommendations in each stage because they are driven by different goals in each stage, i.e., in stage 1 they are comparing alternatives, whereas in stage 2 they are reviewing candidates. For this reason, we are also interested in the moderation role of decision stages, which has not been previously studied.

In the following sections, we firstly introduce the theoretical underpinnings of this research, based on which five hypotheses are proposed. Then we discuss the design of a randomized experiment, which measures consumers' willingness-to-pay across different recommendation scenarios. We present the results of our analysis and discuss the implications for online retail practices and research involving recommender systems. The investigation provides a new angle for understanding the behavioral aspect of recommender systems, as well as guidelines to further improve the design of recommendation agents.

2. THEORETICAL FRAMEWORK

In this section, we discuss the relevant theoretical foundations for our research questions in terms of three dimensions: recommendation type, price, and consumers' decision stage. Based on our primary goal of uncovering the differential effects between different recommendation types, we firstly review research about complementary/substitutable goods in the marketing literature, as well as related empirical studies in the recommender system context. Second, since the basic relationship between complements/substitutes is their cross-price elasticity of demands, we discuss theories describing how one product's price might influence consumers' willingness-to-pay for another product. Finally, we discuss the related literature on consumer decision making processes and how this interacts with the recommender systems.

2.1 Complements and Substitutes in Recommendation

The study of complements and substitutes has long been a central subject in the marketing literature. Generally, products are considered complements (substitutes) if lowering (raising) the price of one product leads to an increase in sales of another [31]. Research shows that consumer choice is easily influenced by context and the set of alternatives available at the time of decision [24], thus, there is significant demand interrelationship among substitutable and complementary goods [20]. Generally, two moderate or strong substitutes should be offered separately, whereas two complements should be offered as a bundle [32], in order to maximize the profit. This is because the introduction of a complement may increase the possibility of buyers finding new uses or added value for existing products [31], whereas the substitutes can be consumed or used in place of one another.

In the online recommendation scenario, a focal product is often presented with several related items as the recommendations. Take Amazon.com as an example, each product is featured on its own designated webpage, along with additional relevant products as recommendations. Hence, a visible directed product network is created whereby products are explicitly connected by hyperlinks [16]. Some studies [22][23][5] have examined the behavioral impacts of recommendation networks, with these studies primarily focusing on the co-purchase recommendation network. Many e-commerce websites provide recommendations from two product networks: co-view and co-purchase product networks. Only recently have their differential effects been considered [16]. More interestingly, co-purchase and co-view networks can be used to implicitly represent two recommendation strategies, that is, recommending complementary and substitutable products, respectively. Although not always the case, it is common that co-purchased items contain complementary products while co-viewed products contain substitutable products. For example, a consumer buy a laptop computer may view several laptops, but purchase only one laptop along with a complementary mouse, software, screen protector or other accessories. Figure 1 and Figure 2 provide an example to this extent with the co-purchase and co-view recommendations on Amazon.com when the focal product is 'Dell Inspiron 15 i5558-5718SLV'.

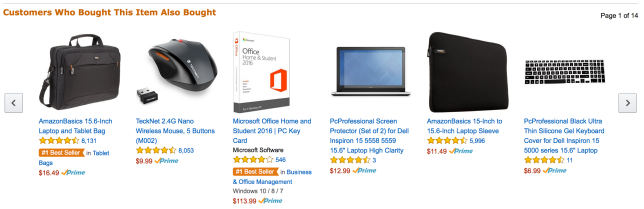


Figure 1. Amazon co-purchase product recommendation

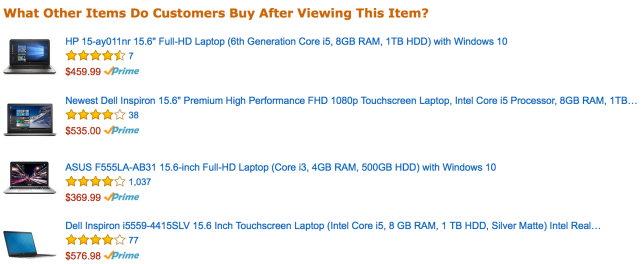


Figure 2. Amazon co-view product recommendation

By definition in microeconomics, if product A and B are complements, increased demand for product A should be associated with increased demand for product B [33]. This complementary product effect leads to the co-purchase network. On the other hand, substitute products have an inverse demand relationship. This leads to the co-view network because consumers tend to view and compare substitutes before making final purchase decisions. Given that recommendations of both complements and substitutes are often presented along with the focal product, it is of significant practical interest to understand their differential effects. Based on economic theory, complements increase demand for the focal product, and the market price will increase accordingly, leading to higher individual willingness-to-pay. Whereas, substitutes decrease demand for the focal product due to competition, which leads to a lower market price and individual willingness-to-pay. Thus, we put forth the following hypothesis:

H1: *Consumers tend to have higher willingness-to-pay for the focal product when it is displayed together with a complementary recommendation as compared to being displayed with a substitutable recommendation.*

Note that this initial hypothesis is intended to test a main effect of recommendation type and is price agnostic. We address the moderating effects of price in the next section.

2.2 Pricing

Researchers in marketing and economics have long recognized that pricing decisions sometimes incorporate more than one product. This is because consumers tend to respond to price relative to some reference price [25], such as the other prices in the store at the point of purchase. Both prospect theory and mental accounting suggest that consumers make decisions based on losses or gains relative to a reference point. When consumers compare the actual price of the focal product with other reference prices, incidental price learning [21] occurs without any explicit intention to memorize them. In offline physical stores, retailers can attempt to influence positively the degree to which the sales of one item affect sales of other items

through in-store product locations and shelf space allocations, for example, locating two complements together. This is very similar to the online recommendations where other products are co-displayed with prices along with the focal product on the webpage. When consumers evaluate the focal product, the prices of recommended products are expected to affect their purchase-related decisions for the focal product. Contrast effect theory [30] suggests that the perceived value of the focal product's price is decreased (increased) when the recommendations presented along with have relatively higher (lower) price. That is, consumers are willing to pay more (less) when the product seems cheaper (more expensive) relative to other products. Thus, we have the following hypothesis about the main effect of *price* on willingness-to-pay:

H2: *The price of a recommended product has a positive influence on consumers' willingness-to-pay for the focal product, that is, consumers are willing to pay more when the price of the recommended product is higher than the price of focal product.*

There are also significant cross-relationships among the sales of substitutable and complementary products [20]. In particular, lower price or promotion of one product can stimulate sales of a complement, whereas supplant sales of other substitutes. That is to say, when the prices of complementary goods go up, the purchase likelihood for the other complementary good may go down, while if the price of one of the substitute goods goes up, the purchase likelihood for others will go up [12]. Furthermore, by influencing the demand of a complementary/ substitutive product through its price, the consumers' willingness-to-pay for that product will be influenced as well [18]. This cross-relationship nature of substitutes and complements indicates that the effect of price is stronger when the competition between two products is high, which is common among substitutes. Therefore, we hypothesize the following:

H2A: *There is an interaction effect between the price and type of recommendations, such that the positive influence of recommendation price on willingness-to-pay for the focal product is stronger when the recommendation type is substitute as compared to complement.*

2.3 Consumers' Two-Stage Decision Making

As illustrated in the previous literature [9][17][29], consumers are often not capable of evaluating all available alternatives in great depth, and this results in a two-stage decision making process. In the first stage, consumers usually browse a large set of available options and identify a small subset of candidates for further consideration. In the second stage, they tend to thoroughly evaluate the candidates and make a final purchase decision. In the second decision stage consumers' motivation and determination to make purchases are increased, thus having higher willingness-to-pay for selected alternatives. Hence, we hypothesize the following main effect of decision stage:

H3: *The decision stage has a positive influence on consumers' willingness-to-pay for the focal product, that is, consumers are willing to pay more when they are in the second stage.*

Given this multistage mental process, Ge et al. (2012) [8] argued that the manner in which information is processed differs systematically between the two decision stages. Their experimental results reveal that the timing of the presentation of specific pieces of information about an alternative across shopping stages has a great impact on consumers' choice. This difference can be attributed to the shopping goals theory [14] and construal level theory [15]. Specifically, consumers are less certain of their shopping goals in the first stage of a shopping process. Thus, their thinking is more abstract when in the first stage. Shopping goals become concrete when they are closer to the final purchase point in the second stage. Therefore, marketing promotions for similar products (i.e., substitutes) are more effective in influencing consumer's spending when their goals are less concrete [10][14]. Researchers have also studied the behavioral effects of recommendations beyond standard substitute recommendations. For instance, Zheng et al. (2009) [33] argued that customers prefer different types of recommendations in different purchase stages. In the first stage of an online purchase process, customers are navigating webpages to compare a large set of similar products. Whereas in the second stage, customers already have a clear candidate set through which to make a purchase decision. In the second stage, substitutive recommendations likely have little impact and recommendations of complement products may be preferred since they introduce items that can add value to the purchase of the focal product. Hence, we hypothesize the following interaction effect:

H3A: *There is an interaction effect between the stage and type of recommendations, such that the stage has a positive effect on willingness-to-pay for the focal product when the recommendation type is complement and it has a negative effect on willingness-to-pay for the focal product when recommendation type is substitute.*

3. EXPERIMENTS

Recommendations on Amazon.com and other platforms generally fall into the complement and substitute product types through the co-purchase and co-view lists. However, this is not always the case, and other contextual factors and user self-selection can impact the effect and content of these recommendations. Therefore, to eliminate these confounding factors and conditions that naturally occur in the field, we designed a randomized controlled experiment so that the recommendation type, recommendation price, and decision stage can be cleanly manipulated. This controlled and randomized treatment approach allowed us to test our hypotheses and make causal inferences.

3.1 Experiment Design and Participants

Our hypotheses express the main effects of each of three main factors (recommendation type, price of recommended product, and decision stage) as well as two two-way interaction effects (price x type and stage x type) on willingness-to-pay for a focal product. Since the focus of the study is on the effects of complementary versus substitutable recommendations, we did not hypothesize the interaction between recommendation price and decision stage. Additionally, since the three-way interaction among these factors is complex and no prior theory provides insights to this regard, this interaction was also not hypothesized.

A factorial experiment was used to test our hypotheses efficiently. Specifically, a 2 (types of recommendation: complements vs. substitutes) x 2 (recommended products' price relative to the focal product's price: low vs. high) x 2 (decision stage: stage 1 vs. stage 2) full-factorial design was used, which results in 8 treatment conditions. Although, the full factorial provides the opportunity to test the three-way interaction and the price x stage interaction, our analysis focuses only on the main effects and interactions identified in our hypotheses. The advantage of factorial experiment designs over randomized controlled trails (RCTs) is that they provide more statistical power with fewer participants. Generally, the objective of RCT is to compare the individual experimental conditions to each other directly, while in a factorial experiment the

combinations of experimental conditions are compared, i.e., the main effects and interactions.

We manipulated the three factors between subjects, who were undergraduates from a business school in a large public university in North America. Subjects received extra credit for their participation in the experiment. We performed a power analysis with the assumption that the effect size of our model will be medium, i.e., $Cohen\ f^2 = 0.15$ (Cohen, 1988). To achieve power (1-β) of 0.80 and a medium effect size, as well as maintain a significance level (α) at 0.05, the minimum sample size for a model with three main effects and two interactions is 92 (the calculation was made by using the package 'pwr' in R).

We published a web link for our online experiment to a large undergraduate class containing approximately 400 students. 261 students clicked on the link to initiate the study. Participants were randomly assigned to one of the eight treatment conditions. The median time of completion is 12 minutes. We dropped observations for 126 participants for the following reasons: not completing the study, completing the experiments in an extremely short time (e.g., less than 4 minutes), completing the study in very long time (e.g., more than 4 hours indicating the study was started, stopped, and started again later), and not passing manipulation checks. Participants were informed that multiple manipulation checks would be used to determine if they took the study seriously, which would then impact whether they received extra credit for their participation. Since we collected the data as an online survey instead of bringing students to a laboratory, it is a common phenomenon that response and completion rates are relatively low [3]. As a result, 135 valid observations were left. The distribution of the valid observations across treatment groups is shown in Table 1. As can be seen in Table 1, the randomly assigned treatments are evenly distributed among participants.

Table 1. Experimental design and sample sizes per group

		complements	substitutes
Stage 1	Low	16	17
Stage 1	High	17	18
Stage 2	Low	17	16
Stage 2	High	17	17

Experiment participants were put in the scenario of purchasing a new computer mouse on an e-commerce site like Amazon.com. We chose a mouse because it is very common for consumers to purchase electronic products online, and a computer mouse has a large number of potential complements and substitutes with both low and high prices.

For the first manipulation factor (i.e., type of recommendation), participants were randomly assigned to one of two different shopping interfaces: the focal product page with recommendations of complementary products, or the focal product page with recommendations of substitute products. The pages included product descriptions directly from Amazon.com. We omitted brand information in the descriptions to eliminate any brand bias. Note that both the complementary and substitutable goods derived from real recommendations from the website. For the second manipulation factor, (i.e., price of the recommended products), participants were randomly assigned to either a high or low price condition. In the high price condition, the recommended products were higher in price than the focal product and in the low price condition the opposite was true. For the third manipulation factor (i.e., decision stage), we designed a two-stage shopping procedure (i.e., consider-then-choose) adapted directly from [8], which the participants followed prior to measuring dependent variables. Participants were randomly assigned to one of two decision stage manipulations: complete the first and second stage of the shopping procedure before being shown the focal product page or complete only the first stage of the shopping procedure before being shown the focal product page. In all treatment groups, the focal product and its posted price and description remained the same.

3.2 Stimuli and Procedures

Participants were first instructed with a cover story that they were participating in research focusing on consumers' preferences and purchase behavior. They were also told that there were no right or wrong answers. Following these initial instructions, participants were randomly assigned to one of the eight treatment groups.

Before the main task of the experiment, participants were asked to answer some basic questions about their opinions on electronic products and were asked to rate several different electronic product categories. Participants were told that their answers to these questions would be used later by our system to predict their preferences and make personalized recommendations for them. This pre-experiment task was used to eliminate their doubt about the basis of recommendations in later steps.

In the main task of the experiment, subjects were asked to shop for a computer mouse and make a purchase decision. We implemented the two-stage shopping decision process based on the methodology used in the marketing literature, (e.g., [8]). In the first stage process, participants were presented with descriptions of 12 computer mice as search results on the e-commerce store. They were asked to browse and evaluate all the product information presented. The 'Next Page' button appeared only after 30 seconds had elapsed, as means of preventing participants from moving ahead too quickly, without reviewing the stage 1 products. Manipulation check questions were also asked to check their impressions about these initial 12 mice. In the second stage, we narrowed down the choice set to 2 mice and participants were asked to evaluate the two items and pick one of them as their final purchase choice. Participants who were randomly assigned to the groups with condition 'stage 1' would only go through the first stage (i.e., browsing information). Comparably, those who were randomly assigned to groups with condition 'stage 2' would go through both the first and the second stage.

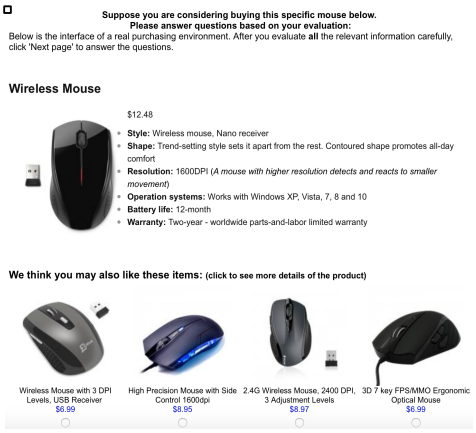


Figure 3. Example Screenshot for the experimental interface (i.e., substitutes recommendation with low price)

After the stage manipulation, participants viewed a specific focal mouse product page. Along with the focal mouse, recommendations were presented according to the random treatment condition. Figure 3 provides example screenshots of the recommendation interface. In the groups with 'low (high) price' condition, all the recommendations' prices were slightly lower (higher) than the price of the focal mouse. Subjects could click on the recommendation to view a detailed description. The number of clicks and duration on each webpage were also recorded.

After viewing the focal product page based on their treatment condition, participants were asked to provide their willingness-to-pay for the focal product. Upon completing the shopping task, participants responded to a set of manipulation check questions and completed a short survey with demographic questions that we use as control variables in our analysis (age, gender, level of education, computer experience, web experience, e-commerce experience, familiarity with and attitudes toward recommender systems).

3.3 Dependent Measure

Willingness-to-pay is the maximum amount an individual is willing to sacrifice to procure a product. Here we adopted the method used by Rucker & Galinsky (2008), Rucker et al. (2014) and Kim & Gal (2014) [13][27][28] to measure willingness-to-pay. Participants indicated their willingness-to-pay using a sliding scale where they could choose from 0% to 120% of the retail price. The interval (i.e., 0%-120%) is used to reduce the amount of response variance and to guard against outliers. We are interested in relative changes in willingness-to-pay due to treatment effects and not in estimating point estimates of willingness-to-pay for specific products, thus the interval willingness-to-pay metric is sufficient. Furthermore, because the market price for the focal product was given, it's not realistic to use the Becker–DeGroot–Marschak approach [7] or second-price auctions to elicit willingness-to-pay. Figure 4 shows the interface for entering willingness-to-pay:

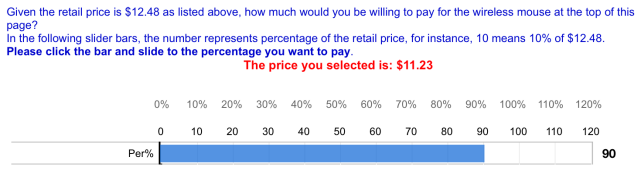


Figure 4. Entering willingness-to-pay

4. RESULTS

Table 2 provides summary statistics on the demographic items collected in our post-experiment survey.

Table 2. Demographic Summary Statistics

Control variables	Summary
Age	Mean: 21.6, SD: 3.35
Gender	50.37% --- female
Primary language	88.89% --- native English speaker
Experience with Internet	88.15% --- spend more than 4 hours per day on the Internet
Experience with e-commerce	77.04% --- browse e-commerce websites more frequently than once a week
Familiarity with RS	85.19% --- familiar with RS
Attitude toward RS	82.22% --- RS is helpful for finding relevant items

4.1 Manipulation Checks

In order to check the saliency of our recommendation type, two questions were asked of participants in the post-experiment survey: *(1) Do you think the products in the section titled 'We think you may also like these items' are complements to the mouse you evaluated?* and *(2) Do you think the products in the section titled 'We think you may also like these items' are substitutes to the mouse you evaluated?*. In terms of the decision stage manipulation check, we did not directly ask subjects' perceptions about *decision stage* because this may be an incomprehensible terminology. Instead, we asked them '*In the previous task you just finished, which procedure(s) have you been through?'*, and provided the following possible responses: (1) *Evaluating a large set of alternative products as if you were gathering information in early stages of shopping* and (2) *Evaluating a small set of alternative products as if you were trying to choose a final one to purchase.* An additional question was used to check participants' perception about the relative price: '*What do you think of the price level of the mouse you just evaluated?'*.

First, to check if participants consciously distinguished between complement and substitute recommendations, we compared their responses toward the two manipulation check questions about recommendation type. They responded with the following 5 claims: "Definitely yes" (coded as 5), "Probably yes" (coded as 4), "Maybe" (coded as 3), "Probably not" (coded as 2), and "Definitely not" (coded as 1). As expected, participants in the complements group perceived the recommendations as complements ($M_{complements} = 4.15, M_{substitutes} = 1.61,\ t(133) = 16.17, p < 0.001$), and not as substitutes ($M_{complements} = 1.42, M_{substitutes} = 4.68, t(133) = -28.87,\ p < 0.001$). The extremely low p-values of these tests help guard against any potential multiple comparison issues. These results support the validity of our manipulation for recommendation types. Further, for the stage check question, participants in different stage conditions correctly perceived their decision stages ($M_{stage1} = 1.19, SD = 0.39, M_{stage2} = 1.97, SD = 0.17, t(133) = -14.81, p < 0.001$). Finally, a successful manipulation check was observed for the price. Due to the contrast effect, people in the high price recommendation condition felt the price of focal product is lower than those assigned in the low price recommendation condition ($M_{high} = 2.96, SD = 0.57, M_{low} = 3.33, SD = 0.68, t(133) = -3.45, p < 0.001$).

4.2 Main Results

Table 3 shows the mean and standard deviation of the willingness-to-pay, measured as a percentage (0%-120%) of the focal product's original price, in each of the eight treatment groups.

Table 3. Mean (SD) Willingness-to-pay (%) in Each Group

Decision stage	*Price*	*complements*	*substitutes*
Stage 1	Low	68.19 (15.86)	79.29 (17.66)
Stage 1	High	87.65 (13.50)	93.78 (21.81)
Stage 2	Low	88.29 (14.29)	81.63 (16.04)
Stage 2	High	93.94 (10.87)	89.53 (13.14)

To test the proposed hypotheses, we need to make comparisons between combinations of groups – the main and interaction effects. Since there are three manipulated factors in the experiment, we started by conducting a three-factor Analysis of Variance (ANOVA) and the results are presented in Table 4. The results reveal that the main effect of *stage* and *price*, as well as the interaction between *recommendation type* and *stage* are significant. We also conducted orthogonal contrast analysis, which provided consistent results and is omitted due to space constraints. Details can be obtained by contacting the authors directly.

Table 4. Results of Three-factor ANOVA

	Df	SSE	MSE	F value	Pr(>F)
Type	1	73	73	0.279	0.5981
Price	1	4670	4670	17.786	4.66e-05 ***
Stage	1	1200	1200	4.570	0.0345 *
Type : Price	1	9	9	0.036	0.8505
Type : Stage	1	1688	1688	6.430	0.0124 *
Price : Stage	1	872	872	3.321	0.0708 +
Residuals	127	33343	263		

Significance levels: + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Figure 5 displays the average willingness-to-pay under the combined conditions. Specifically, there is no significant difference between groups with complement and groups with substitute recommendations ($M_{complements} = 84.76,\ SD = 16.67, M_{substitutes} = 86.24, SD = 18.85, F = 0.279, p = 0.598$). However, when the prices of recommended products are relatively high, participants' willingness-to-pay is much higher than that in relatively low prices condition($M_{low} = 79.48,\ SD = 17.48, M_{high} = 91.26, SD = 15.75, F = 17.786, p < 0.001$), which supported H2. Similarly, the difference between conditions in stage 1 and stage 2 was in the expected directions ($M_{stage1} = 82.60, SD = 19.99, M_{high} = 88.45, SD = 14.26, F = 4.570, p < 0.05$), thus supporting the hypothesis that consumers are willing to pay more when they are in the second decision stage (H3). Further, we calculated Cohen's *d* to capture the effect size. Cohen's *d* is known as the difference of two population means and divided by the standard deviation from the data. The effect size for *price* and *stage* factors are 0.669 and 0.372, which indicate medium-to-large and small-to-medium effect, respectively.

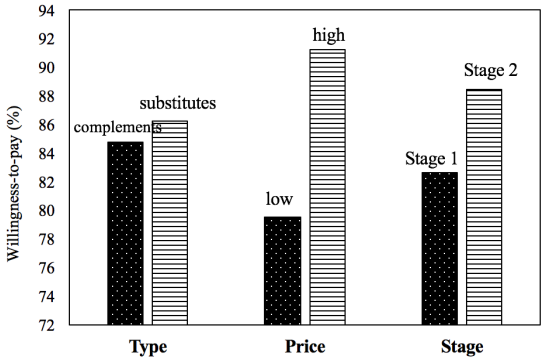


Figure 5. Average Willingness-to-pay in Combined Conditions

For the interaction effects, Figure 6 demonstrates the difference of mean values for complements and substitutes groups under different price levels and decision stages, respectively. The left figure shows no interaction between *type* and *price*, since both complements and substitutes groups have higher willingness-to-pay in high relative price conditions ($F = 0.036, p = 0.851$). The crossing lines in right figure indicate significant interaction effect between *type* and *stage* ($F = 6.430, p < 0.05$). Particularly, the effect of *decision stage* is much stronger when the recommendations are complements ($M_{stage1} = 78.21, SD = 17.62, M_{stage2} = 91.12, SD = 12.28, t(65) = -3.29, p < 0.001$), while it's not significant under substitute conditions ($M_{stage1} = 86.74, SD = 21.18, M_{stage2} = 85.70, SD = 15.14, t(66) = 0.23, p = 0.41$). Therefore, our hypothesis of interaction effect H3A is partially supported, and H2A is not supported.

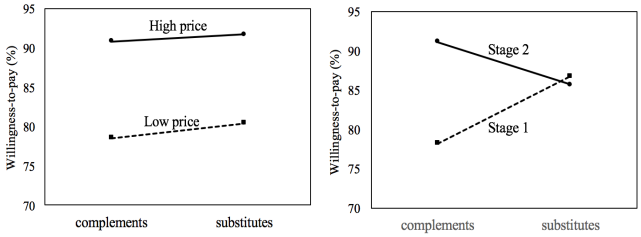


Figure 6. Interaction effects (Left: type × price; Right: type × stage)

Furthermore, to get the effect size of our model as well as coefficients of each factor, we estimated a sequential linear model. Firstly, we regressed the willingness-to-pay on a set of control variables, including gender, preference to the focal product, experience with e-commerce, familiarity with and attitude toward recommender systems. After that, we included the five independent variables of interests to the model. The *type* factor has two levels, either complements (0) or substitutes (1), the *price* factor is either low (0) or high (1), and the *stage* factor is either stage 1 (0) or stage 2 (1). The regression results of these two models are shown in Table 5, and the R-square increased 0.149 after including these five variables. Consistent with the ANOVA results, we got significant positive coefficients for *price* and *stage*, indicating that consumers have higher willingness-to-pay in the high price condition (compared to low price) and stage 2 condition (compared to stage 1). In addition, the interaction effect between type and stage is also marginally significant at level α = 0.1. The coefficients in the table suggest that consumers shown a recommended product with high price reported 10.353% higher willingness-to-pay in terms of the retail price. Similarly, consumers in the second stage reported 10.832% higher willingness-to-pay in terms of the retail price than those in the first stage.

Table 5. Results of the Linear Regression Models

	Dependent variable: Willingness-to-pay (%)					
	Model 1: Control ($R^2 = 0.1699$)			Model 2: Full model ($R^2 = 0.3189$)		
	Coefficient (SE)	T statistic	P-value	Coefficient (SE)	T statistic	P-value
Intercept	53.166 (10.450)	5.088	1.25e-06***	44.589 (9.862)	4.521	1.42e-05***
Type				5.572 (4.828)	1.154	0.251
Price				10.353 (3.769)	2.747	**0.007** **
Stage				10.832 (4.008)	2.703	**0.008** **
Type * Price				2.279 (5.336)	0.427	0.670
Type * Stage				-10.729 (5.691)	-1.885	**0.062** +
Preference	7.141 (1.539)	4.640	8.45e-06***	6.232 (1.492)	4.178	5.50e-05***
Gender	3.255 (3.081)	1.056	0.292	3.803 (2.905)	1.309	0.193
Experience	-1.126 (1.616)	-0.697	0.487	-1.054 (1.527)	-0.690	0.492
Familiarity	0.600 (1.548)	0.388	0.699	1.412 (1.491)	0.947	0.345
Attitude	-6.362 (3.032)	-2.098	0.0378 *	-6.924 (2.807)	-2.467	0.015 *

Significance levels: + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

The effect size of our sequential multiple regression model is calculated by Cohen's f^2. It is defined as $f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$, where R^2_A is the variance accounted for by a set of control variables A and R^2_{AB}

is the combined variance accounted by A and another set of independent variables of interest B. Here in our model, we have $R_A^2 = 0.1699$ and $R_{AB}^2 = 0.3189$, resulting in a medium-to-large effect size of $f_B^2 = 0.219$. We also conducted a post-hoc power analysis and with the 135 observations and the calculated effect size, the power of our model is 0.993 while maintaining the significance level at 0.05. This provides evidence that the null effects are true and not the result of a lack of power.

Since we measured the willingness-to-pay by restricting participants' choice from 0% to 120% of the stated retail price, it may result in censored and non-normal data. Therefore, we performed robustness check that removed the normality assumption. We conducted two non-parametric tests and ran a Tobit regression model using both dummy coding (0,1) and effect coding (-1,1). The coefficients and significance levels are consistent with our baseline analysis. Due to space constraints, details of the robustness check are omitted.

5. DISCUSSION AND CONCLUSIONS

5.1 Summary of Findings

In this paper, we conducted a randomized experiment to examine the impact of complement and substitute recommendations on consumers' willingness-to-pay for the focal product. Table 6 summarizes our findings which corresponds to the proposed hypotheses.

Table 6. Hypotheses and Results

Hypotheses	Results
H1: Consumers tend to have higher willingness-to-pay about the focal product when it is displayed together with a complementary recommendation as compared to being displayed with a substitutable recommendation.	Not Supported
H2: The price of a recommended product has a positive influence on consumers' willingness-to-pay for the focal product, that is, consumers are willing to pay more when the price of the recommended product is higher than the price of focal product.	**Supported**
H2A: There is an interaction effect between the *price* and *type* of recommendations, such that the positive influence of recommendation price on willingness-to-pay for the focal product is stronger when the recommendation type is substitute as compared to complement.	Not supported
H3: The decision stage has a positive influence on consumers' willingness-to-pay for the focal product, that is, consumers are willing to pay more when they are in the second stage.	**Supported**
H3A: There is an interaction effect between the *stage* and *type* of recommendations, such that the *stage* has a positive effect on willingness-to-pay for the focal product when the recommendation type is complement and it has a negative effect on willingness-to-pay for the focal product when recommendation type is substitute.	**Partially Supported**

Experimental results provide evidence that there is no significant main effect difference between complementary and substitutable recommendations on consumers' willingness-to-pay for the focal product. We further investigated two factors that commonly present with recommendations: decision stage and the price of recommended products. We found that consumers are willing to pay more for a specific product as decision stage increases. An interesting finding is the interaction between recommendation type and decision stage. The positive effect of stage vanished when the recommendation is substitute to the focal product, while it is very significant with complementary recommendations. This is consistent with previous findings that customers prefer different recommendations against different purchase stages, as well as highlighting the importance of timing in recommender systems. The price of recommended products was also found to have significant effects on willingness-to-pay. Serving as a reference point, the prices of recommendations may be compared with the retail price of the focal product, which could cause consumers to adjust their willingness-to-pay through incidental price learning. Under the condition with high recommendation prices, consumers tend to have higher willingness-to-pay for the focal product and vice versa. This positive effect is significant no matter the type of recommendation for other products.

5.2 Theoretical Contributions

Our research offers important theoretical contributions in the following ways. First, studies on product recommendations have focused on the consumers' different preferences and behaviors for one products in the presence of recommendations [1][2]. This paper extends the behavioral research on recommender systems by studying the question whether recommending 'other products' on the same webpage had an effect on consumers' willingness-to-pay for the focal product. Second, prior research has not paid much attention to different types of recommendations. Deriving from economics literature, two typical relationships between products are examined, that is, complements and substitutes. Third, our research is one of the few studies that examine the detailed recommendation features, i.e., price of recommended products as well as consumers' decision stage. Integrating the consumers' decision process, we have a better understanding of the behavioral aspects of recommendations in online purchases.

5.3 Implications for Practice

Beyond contributing to advancing the academic literature, our findings also have significant practical implications and may guide the platform's recommendation strategy. The vulnerability of consumers' willingness-to-pay indicates the importance of 'other products' in recommender systems. This suggests new possibilities for influencing product sales by manipulating the contextual information of recommendations. Another important implication is about the timing of recommendations, i.e., complementary recommendations should be delayed to the second decision stage.

5.4 Future Work

The main limitation of this study is that we are not observing real world purchases. In contrast, however, an advantage is that our controlled randomized experiment allows us to make causal inferences, and thus trading external validity for identification. Future research can be developed by exploring other factors associated with recommended products, such as average ratings, quality, pictures of complements/substitutes and so on. Additionally, we can use observational data to empirically validate the findings of our experiment. By examining the relationships between recommendation network properties and products' sales, we will have additional support for the influence of complementary and substitutable product recommendations on consumers' economic behavior from an aggregate level.

6. REFERENCES

[1] Adomavicius, G., Bockstedt, J.C., Curley, S.P. and Zhang, J. 2014. Suggest or Sway? Effects of Online Recommendations on Consumers' Willingness to Pay. (2014). *Working paper*.

[2] Adomavicius, G., Bockstedt, J.C., Curley, S.P. and Zhang, J. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*. 24, 4 (2013), 956–975.

[3] Baruch, Y. and Holtom, B.C. 2008. Survey response rate levels and trends in organizational research. *Human Relations*. 61, 8 (2008), 1139–1160.

[4] Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

[5] Dhar, V., Geva, T., Oestreicher-singer, G. and Sundararajan, A. 2014. Prediction in Economic Networks. *Information Systems Research*. 25, 2 (2014), 264–284.

[6] Donaldson, C., Jones, A.M., Mapp, T.J. and Olson, J.A. 1998. Limited dependent variables in willingness to pay studies: applications in health care. *Applied Economics*. 30, 5 (1998), 667–677.

[7] Frederick, S. 2012. Overestimating Others' Willingness to Pay. *Journal of Consumer Research*. 39, 1 (2012), 1–21.

[8] Ge, X., Häubl, G. and Elrod, T. 2012. What to Say When: Influencing Consumer Choice by Delaying the Presentation of Favorable Information. *Journal of Consumer Research*. 38, 6 (2012), 1004–1021.

[9] Häubl, G. and Trifts, V. 2000. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*. 19, 1 (2000), 4–21.

[10] Ho, S.Y., Bodoff, D. and Tam, K.Y. 2011. Timing of Adaptive Web Personalization and Its Effects on Online Consumer Behavior. *Information Systems Research*. 22, 3 (2011), 660–679.

[11] Hosanagar, K., Fleder, D., Lee, D. and Buja, A. 2014. Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. *Management Science*. 60, 4 (2014), 805–823.

[12] Jin, R.K.-X., Parkes, D.C. and Wolfe, P.J. 2007. Analysis of bidding networks in eBay: Aggregate preference identification through community detection. *AAAI Workshop - Technical Report*. WS-07-09, (2007), 66–73.

[13] Kim, S. and Gal, D. 2014. From Compensatory Consumption to Adaptive Consumption: The Role of Self- Acceptance in Resolving Self-Deficits. *Journal of Consumer Research*. 41, August (2014), 526–542.

[14] Lee, L. and Ariely, D. 2006. Shopping goals, goal concreteness, and Conditional Promotions. *Journal of Consumer Research*. 33, June (2006), 60–71.

[15] Liberman, N., Trope, Y. and Wakslak, C. 2007. Construal level theory and consumer behavior. *Journal of Consumer Psychology*. 17, 2 (2007), 113–117.

[16] Lin, Z., Goh, K.-Y. and Heng, C.-S. 2015. The demand effects of product recommendation networks: an empirical analysis of network diversity and stability. *MIS Quarterly*. (2015). *Forthcoming*.

[17] Liu, Q. and Arora, N. 2011. Efficient Choice Designs for a Consider-Then-Choose Model. *Marketing Science*. 30, 2 (2011), 321–338.

[18] Loomis, J., Gonzalez-Caban, A. and Gregory, R. 1994. Do Reminders of Substitutes Influence Contingent Valuation Estimates? *Land Economics*. 70, 4 (1994), 499–506.

[19] McAuley, J., Pandey, R. and Leskovec, J. 2015. Inferring Networks of Substitutable and Complementary Products. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015).

[20] Mulhern, F.J. and Leone, R.P. 1991. Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store Profitability. *Journal of Marketing*. 55, 4 (1991), 63–76.

[21] Nunes, J.C. and Boatwright, P. 2004. Incidental Prices and Their Effect on Willingness to Pay. *Journal of Marketing Research*. 41, 4 (2004), 457–466.

[22] Oestreicher-Singer, G. and Sundararajan, A. 2012a. Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*. 36, 1 (2012), 65–83.

[23] Oestreicher-Singer, G. and Sundararajan, A. 2012b. The visible hand? Demand effects of recommendation networks in electronic markets. *Management Science*. 58, 11 (2012), 1963–1981.

[24] Payne, J.W., Bettman, J.R. and Johnson, E.J. 1992. Behavioral decision research: a constructive processing perspective. *Annual Reviews of Psychology*. 43, 1 (1992), 87–131.

[25] Rajendran, K.N. and Tellis, G.J. 1994. Contextual and temporal components of reference price. *Journal of Marketing*. 58, 1 (1994), 22–34.

[26] Rao, A.R. and Sieben, W. a. 1992. The Effect of Prior Knowledge on Price Acceptability and the Type of Information Examined. *Journal of Consumer Research*. 19, 2 (1992), 256–270.

[27] Rucker, D.D. and Galinsky, A.D. 2008. Desire to Acquire: Powerlessness and Compensatory Consumption. *Journal of Consumer Research*. 35, 2 (2008), 257–267.

[28] Rucker, D.D., Hu, M. and Galinsky, A.D. 2014. The Experience versus the Expectations of Power: A Recipe for Altering the Effects of Power on Behavior. *Journal of Consumer Research*. 41, August (2014), 381–396.

[29] Russo, J.E. and Leclerc, F. 1994. An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*. 21, September (1994), 274–290.

[30] Sherif, M. and Hovland, C.I. 1961. Social judgment: Assimilation and contrast effects in communication and attitude change. (1961).

[31] Shocker, A., Bayus, B. and Kim, N. 2004. Product Complements and Substitutes in the Real World: The Relevance of "Other Products." *Journal of Marketing*. 68, 1 (2004), 28–40.

[32] Venkatesh, R. and Kamakura, W. 2003. Optimal Bundling and Pricing under a Monopoly: Contrasting Complements and Substitutes from Independently Valued Products. *The Journal of Business*. 76, 2 (2003), 211–231.

[33] Zheng, J., Wu, X., Niu, J. and Bolivar, A. 2009. Substitutes or Complements: Another Step Forward in Recommendations. *Proceedings of the 10th ACM conference on Electronic commerce*. (2009), 139–145.

[34] Zhu, T., Harrington, P., Li, J. and Tang, L. 2014. Bundle recommendation in ecommerce. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*. (2014), 657–666.

DiRec: A Distributed User Interface Video Recommender

Wessam Abdrabo
Technical University of Munich
Boltzmannstrasse 3
85748 Garching Bei München, Germany
wessam.abdrabo@in.tum.de

Wolfgang Wörndl
Technical University of Munich
Boltzmannstrasse 3
85748 Garching Bei München, Germany
woerndl@in.tum.de

ABSTRACT

Distributed User Interfaces (DUIs) are graphical interfaces whose components are distributed in one or many of the UI distribution dimensions: Time, space, platforms, displays, or users. In this work, we have investigated the impact of the application of DUIs, with respect to the different DUI dimensions, on the experience of users of recommender systems. We developed two prototype video recommendation mobile applications: Monolithic Interface Recommender (MiRec), and Distributed Interface Recommender (DiRec). Sharing mostly the same interface, DiRec additionally offers the possibility of migrating parts of the UI between the mobile application and a larger display (LD). A user study was conducted in which participants used and evaluated both MiRec and DiRec. Our results show a significant difference between DiRec and MiRec in attractiveness (general impression and likability), stimulation, and novelty measures, which posits the existence of a strong interest in DUI recommender systems. Nonetheless, MiRec was found more easy-to-learn and easier to understand than DiRec which gives room for further investigation to pinpoint the reasons of DiRec's relatively lower perspicuity measures.

CCS Concepts

•**Human-centered computing** → **User interface design;**

Keywords

Distributed User Interfaces; Recommender Systems; Migratable Interfaces; Mobility; User Study.

1. INTRODUCTION

With the advancement of ubiquitous computing and the trend of the ever-increasing number of devices per user, users of interactive systems no longer perform tasks that reside mainly on a single device, but are rather confronted with situations where they need to complete tasks across several platforms. A typical situation is a user carrying out tasks in a multi-device environment that presents itself effectively to the user as a single UI, but which is actually distributed along these platforms. Such situations represent typical cases of Distributed User Interfaces (DUIs). Hence, DUIs represent an attempt to overcome the limitations of user interfaces that are manipulated by a single user, on a single platform, in a fixed environment, providing few or no variations along these distribution dimensions.

To our best knowledge, surveyed studies for the applications of DUIs do not include any which tackle single-user recommender systems; the fact that provided the main motivation for this research. We hypothesize that the distribution of recommender systems' UIs leads to an enhanced user experience. To verify our hypothesis, we developed two high fidelity prototypes for video recommendation: Monolithic Interface Recommender (MiRec), which is a conventional mobile video recommendation application, and Distributed Interface Recommender (DiRec), which is a distributed version of the mobile video recommender where the interface is distributed among a mobile device (SD) and a large-display screen (LD).

The proceeding sections describe this research's main contributions: A proposal for a generic model for UI distribution for recommendation applications, the design of DiRec which is considered as an instance of this generic model, as well as the results and conclusion of a user study that was conducted to test the impact of our DUI recommender's design on users' experience.

2. BACKGROUND AND RELATED WORK

Enhancing the experience of users of recommender systems through developing more sophisticated recommendation algorithms, taking in consideration aspects such as the novelty, diversity, and accuracy of recommendations, has become the focus of many recent studies. However, fewer studies investigate the possibility of enhancing the user's experience through providing novel UI solutions for recommenders. None of the surveyed research has considered the impact of the distribution of the UI of recommenders on the user's experience. This is where our study provides its main contribution.

During the course of our investigation, we surveyed many studies that laid the foundation of the relatively new field of DUIs. Mostly relevant to our study is Vanderdonckt et al. [9] 's description of what constitutes a distributed UI environment: "UI distribution concerns the repartition of one or many elements from one or many user interfaces in

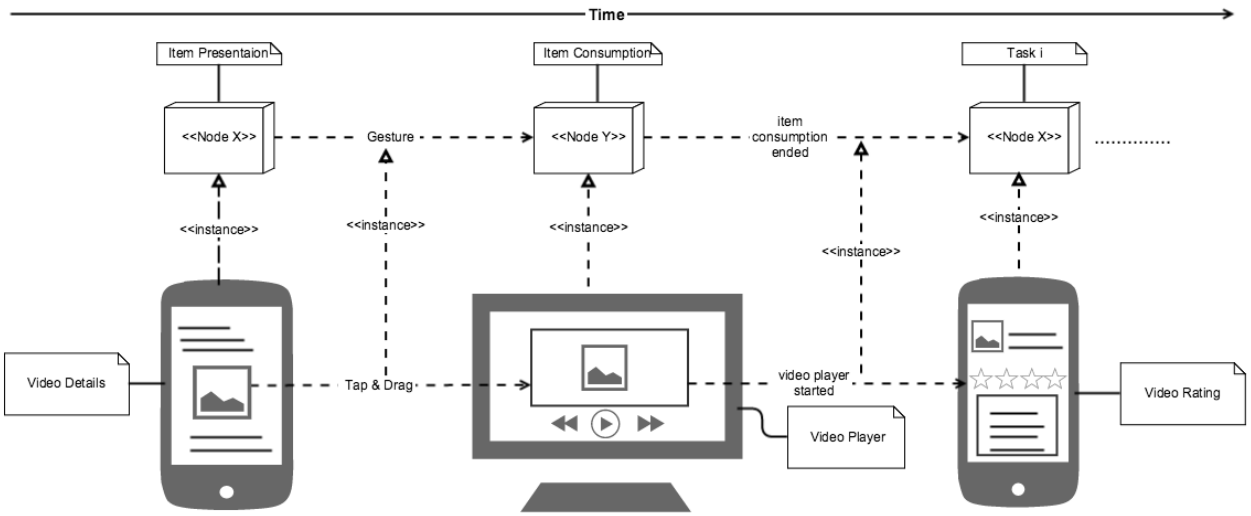


Figure 1: Recommended video consumption and rating as an instance of the generic DUI model.

order to support one or many users to carry out one or many tasks on one or many domains in one or many contexts of use, each context of use consisting of users, platforms, and environments." To deepen our understanding of the various dimensions of UI distribution, we surveyed several studies ([2], [3], [5], [9]). However, one that has been especially relevant to our study is the 4C model described by Demeure et al., through which we could define the 4Cs of our proposed DUI recommender: Computation (what is distributed?), in other words the element of distribution, which could be the task or the platform, Communication (when is it distributed?) or time, Coordination (who initiates distribution?) which is a variation on the user dimension, and Configuration (from where and to where is the distribution operated? on the physical pixel level, or the logical level) [2].

On the other hand, a number of studies have found DUI techniques useful for their applications among which are IAM [1], Aura [7] and ConnecTables [8].

For implementation of our DUI recommender, we adopt a dual display (SD-LD) approach which is similar to Kaviani et al's, who argue that the use of ubiquitous cell phones as an SD component in a DUI not only offer a means to interact with LD displays, but increasingly offer a small, but high quality screen to complement the LD [4].

Moreover, in our previous work [10], we investigated the application of DUIs in group recommender systems. We developed a scenario of a movie recommender, where the UI is distributed among two platforms: a PDA that works as a small display (SD) and a table-top that works as a large display (LD). Users get to view and rate recommended items on their PDAs individually, and as a group, they get to reach a consensus by doing the voting on the table-top. This DUI solution to the voting part of group recommendation is proved by the study to improve the process of reaching consensus among a group. This study takes a further step by investigating the benefits of using DUIs in single-user recommender systems.

3. DESIGN OF A DUI SINGLE-USER RECOMMENDER

Scenarios of our DUI video recommender depict a multi-device environment, in which the flow of control (logic) and the application's user interface are decoupled in a way that allows for the distribution of UI components along the different devices. In other words, the user of such a system is provided with a distributed solution, which enables him/her to perform tasks on whichever device in this environment (by for example migrating the UI components between the different devices) independently of where the application is running, and of the constraints presented by the different platforms running the application.

3.1 Generic Model for UI Distribution

The following are generic scenarios for UI distribution of interactive systems that are applicable to recommender systems:

- *Migration of Item Consumption*: present the recommended content on one device while giving the user the ability to consume the content on another device.
- *Performing Parallel Activities*: user can perform tasks simultaneously and independently from each other.
- *Overview and Detail Presentations*: show different versions of the presented content at different levels of granularity on different nodes.
- *Content Filtering*: distribute the task to filter the user's choice of what to consume.
- *Content Redirection*: content could be transferred to be presented on a different node.
- *Migration of Items Between Users*: content redirection/migration of a list of recommended items (or an item in this list) from one user of the system to one or more other users.

We will describe more specific scenarios that can be considered as an extension of this generic UI distribution model (Figure 1) in a distributed video recommender application in the next subsection.

3.2 DiRec: Distributed Interface Video Recommender

We assume the users are working with a smaller (SD), e.g. a smartphone or other mobile device, and a larger display (LD), e.g. a display screen.

3.2.1 *Pre-Configuring UI Distribution Options*

This scenario presents the initiation point of the system, in which the user is given an option to pre-configure the different options the system offers for UI distribution, and hence be the initiator of UI distribution. This offers the ability to delay the decision of which UI components to present on which platform, making the system distributed in time. This is made possible by presenting the user with a Meta UI in which he/she is asked to drag and drop the components of their choice to the target platform.

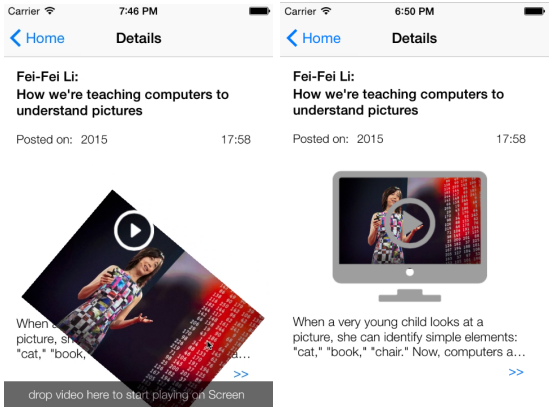


Figure 2: Redirecting recommended item consumption from SD to LD.

3.2.2 *Presentation of Recommendation Results*

The presentation of recommended videos is shown in parallel on the SD and LD, however, in different levels of granularity. The mobile device shows a detailed list of all the recommended videos, together with detailed information about the video, in tabular form with different categorizations. On the LD, an overview presentation is shown for the recommended items that scored the highest for the user without details, however shown in different sizes to indicate the recommendation scores.

3.2.3 *Recommended Item Details Presentation*

Moreover, in our proposed design, we offer the possibility of distributing parts of the UI with a fine granularity. The user selects a single table-cell in the videos list and could move it to the LD by applying the gesture, as opposed to just mirroring or transferring the UI at a more coarse granularity.

3.2.4 *Recommended Item Consumption and Rating*

Starting a video on the LD is done as depicted in Figure 2 in our prototype. On the video details page on the mobile device (SD), the user performs a pan gesture on the video image, which then triggers the migration of the video consumption from the mobile device to the LD.
The video player automatically starts on the LD, providing the user with all controls for the video playback. After the video playback starts automatically on the LD, the LD triggers the mobile device to display the rating page for the user on the SD. Hence, the two tasks could be carried out simultaneously by the user (Figure 3).

3.2.5 *Filtering Recommended Items*

Filtering is done by performing a right swipe gesture on the video item in the list on the SD which redirects the content of the video to the LD. The display of the content on the LD is also done in an overview-detail coupling manner. After the user is done filtering the LD will contain all the selected items displayed as an overview.

3.2.6 *Redirecting Favorites Lists*

Unlike previously described scenarios which involve a single user of the system, this scenario involves two or more users. On the SD, the user selects a favorite-items list. On applying a long-press on the list, the user is prompted with a list of users from which he could select one or more users to transfer this list to.

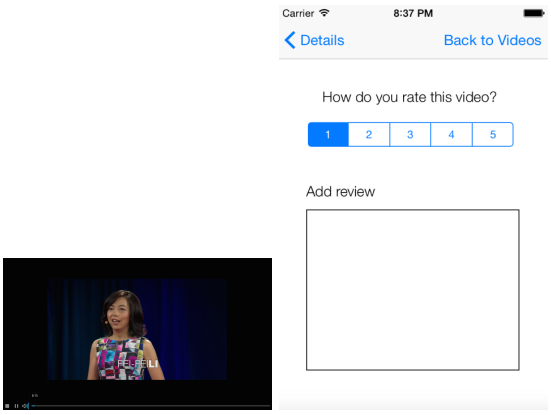


Figure 3: Rating a recommended video on SD in parallel to watching it on LD.

3.3 Prototype Implementation

A subset of the suggested distribution scenarios was selected for implementation. MiRec is developed as the non-distributed version of DiRec and is meant for comparison with DiRec's interface through our comparative user study. Both applications share mostly the same design, however, thorough DiRec, the user could complete tasks in a distributed manner between a mobile application and a large display screen, while with MiRec, users could only complete tasks on the mobile device. MiRec is developed as an iOS mobile application while DiRec is distributed along an iOS application and an LD Python application with a communication layer in between which mainly relies on light-weight TCP-IP based message passing between both platforms (e.g.: play:<videoID> is passed from SD to LD in DiRec to play a video on LD).

4. USER STUDY

To evaluate our approach, we have conducted a user study in three phases. 24 participants were asked to use both MiRec and DiRec and rate their experiences of the products using the User Experience Questionnaire (UEQ) method [6] shortly after finishing the test.

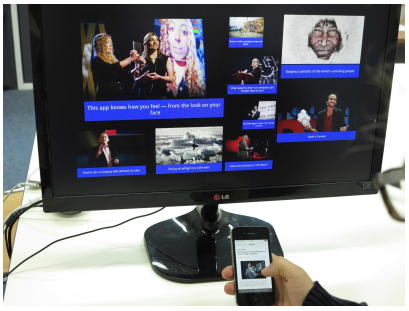


Figure 4: Participant's interaction with DiRec.

4.1 Setup

Each participant was first briefed about how to use MiRec and DiRec, then he/she was asked to complete a set of tasks on both applications including navigating recommendations' lists, playing and rating of videos. Each participant was given an iPhone with both DiRec and MiRec installed and was being asked to interact with the LD screen component during the course of the experiment (Figure 4). During the last phase of the experiment, participants were asked to give their direct impression of the application using the UEQ method [6]. UEQ consists of 6 scales with 26 items which measure Attractiveness (overall impression or the likability), Perspicuity (learnability and ease-of-use), Efficiency (the ability to perform tasks without exerting extra effort), Dependability (user's control over the experience), Stimulation (excitement and motivation) and Novelty (innovation and creativity).

4.2 Results

Figure 5 shows the result of UEQ's comparison of MiRec (left side, blue) and DiRec (right side, red). With respect to attractiveness, stimulation, and novelty, DiRec scores higher than MiRec. For efficiency and dependability, they measure almost similarly with MiRec scoring slightly better than DiRec. MiRec, however, scores much higher than DiRec when it comes to the perspicuity scale. Conducted t-Tests showed statistical significance with regard to perspicuity (α =0.0092), stimulation (α=0.0007), and novelty (α =0.0000), but no significance for attractiveness, efficiency and dependability with an alpha level of 0.05.

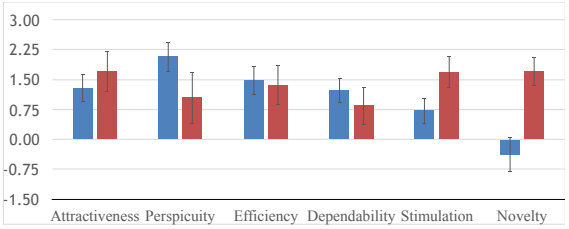


Figure 5: Comparison of scale means in MiRec (left/blue) and DiRec (right/red)

5. CONCLUSIONS AND FUTURE WORK

This work investigates the impact of using distributed user interfaces on the experience of users of recommendation applications. Our comparative user study's UEQ results could be interpreted as follows: The use of DUIs aids the stimulation and novelty of recommendation applications, hence, enriches the user's experience, does not hinder the efficiency or limit the span of the user's control of recommendation applications, results in more attractive recommendation applications, however, might affect the learnability and ease-of-use of recommendation applications. Notwithstanding the promising results of our study, the study has fallen short in providing an explanation of whether the relatively lower perspicuity measures of DiRec is a result of insufficient explanation of the study's procedure, or if it was DiRec's design that was relatively less easy to understand and learn. A possible future work would be to further investigate this aspect. Lastly, we strongly believe that giving more span of control to the user through allowing pre-configuration of UI distribution schemes could further enhance the DUI experience.

6. REFERENCES

[1] J. Coutaz, L. Balme, C. Lachenal, and N. Barralon. Software infrastructure for distributed migratable user interfaces. In *Proc. of UbiHCISys Workshop on UbiComp*, volume 2003. Citeseer, 2003.

[2] A. Demeure, J.-S. Sottet, G. Calvary, J. Coutaz, V. Ganneau, and J. Vanderdonckt. The 4c reference model for distributed user interfaces. In *Autonomic and Autonomous Systems, 2008. ICAS 2008. Fourth International Conference on*, pages 61–69. IEEE, 2008.

[3] N. Elmqvist. Distributed user interfaces: State of the art. In *Distributed User Interfaces*, pages 1–12. Springer, 2011.

[4] N. Kaviani, M. Finke, R. Lea, and S. Fels. Dual displays: towards an interaction model and associated design guidelines. *DUI 2011*, page 69, 2011.

[5] J. Melchior. Distributed user interfaces in space and time. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, pages 311–314. ACM, 2011.

[6] M. Schrepp. *User experience questionnaire handbook*. ueq-online.org, 2015.

[7] J. P. Sousa and D. Garlan. Aura: an architectural framework for user mobility in ubiquitous computing environments. In *Software Architecture*, pages 29–43. Springer, 2002.

[8] P. Tandler, T. Prante, C. Müller-Tomfelde, N. Streitz, and R. Steinmetz. Connectables: dynamic coupling of displays for the flexible creation of shared workspaces. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 11–20. ACM, 2001.

[9] J. Vanderdonckt et al. Distributed user interfaces: how to distribute user interface elements across users, platforms, and environments. *Proc. of XI Interacción*, 20, 2010.

[10] W. Wörndl and P. Saelim. Voting operations for a group recommender system in a distributed user interface environment. In *RecSys Posters*. Citeseer, 2014.

Learning User's Preferred Household Organization via Collaborative Filtering Methods

Stephen Brawner
Brown University
Providence, RI
brawner@cs.brown.edu

Michael L. Littman
Brown University
Providence, RI
mlittman@cs.brown.edu

ABSTRACT

As learning robots and smart devices become common household occurrences, their users will be required to invest more time to train them on the details specific to their household and lifestyle. This *burden of personalization* may eventually become a roadblock to the adoption of smart devices and robots. We are interested in reducing the burden of personalization by leveraging learned information from other households. However, machine learning methods incorporating such data will require smart recontextualizations that can map the preferences from a collection of similar users onto the user's own household space. We present several collaborative filtering based methods to solve the problem of a robot organizing the items in a kitchen for their user: a traditional collaborative filtering method based on prior work incorporating user's item–item distance ratings, and a context-aware collaborative filtering method, which enables direct learning of item–location ratings. We present results on user-annotated kitchens.

CCS Concepts

•Information systems → Recommender systems; •Applied computing → *Command and control;*

Keywords

applications, robotics, context-aware, collaborative-filtering

1. INTRODUCTION

As learning robots become common household items, people will be required to train their robot on the details specific to their household and lifestyle. In the context of household automation, smart devices like the Nest thermostat can learn user's preferred heating and cooling schedules from tracked manual inputs [11]. Even more so than smart devices, robots will need to learn significantly larger amounts of household information, putting a substantial burden on the user for training. As this *burden of personalization* becomes more commonplace in home automation and robotics, users may find themselves investing more and more time into training their smart devices and robots. We are interested in reducing the burden of personalization by leveraging learned information from other households.

The challenge lies in the inability to directly transfer learned information from one household to another, due to uniqueness of household designs, schedule, demographics and composition of the household members, and the combined preferences of all users interacting with the system. Therefore, it is important to identify certain contextual facets of the learning problem to work on.

We consider one such personalization problem in the area of household robotics: the problem of a robot tasked with learning the organization of items in a user's kitchen like putting away groceries or the items from a dishwasher. Asking for the location of every item would be time consuming—possibly more so than having the person put away the items themselves. Therefore, we were motivated to design an algorithm that leverages the organization of other users' kitchens to help predict the user's organization of items.

In the context of videos, products, and news articles, collaborative filtering techniques used for recommender systems have been developed to transfer the experience of other users to help identify items a target user would find desirable. Typically, these techniques predict ratings of user-item pairs, that is, how would a given user rate an item based on their past ratings [18]. Recently, *context* has been incorporated into recommender systems to improve their performance—a crucial element for our application.

Context-aware collaborative filtering, for example as presented by Panniello et al. [12] and Rendle et al. [17], assumes that more information exists that predicts or informs a user's choice of ratings than just the item itself. Context may include information about the user or about the item being rated. In the case of household organization, the context can include the set of possible locations, the identities of the items known to be in those locations, or categorical information related to the items and locations. For this application, we envision a robot with sufficient perceptual abilities to identify these pieces of context.

One of our proposed methods employs a context-based recommender system, factorization machines (FMs) [15], to predict a user's item–location ratings—the degree of acceptability of the placement of a given item at a given location in a kitchen. In contrast, we also present an alternative collaborative filtering method that predicts locations based on

the item's predicted rating distance to items already placed in that location, which builds directly on prior work [1]. The important difference is, while the previous collaborative filtering technique required explicit item–item ratings, the FM method enables direct predictions of item–location ratings, which can even be boosted by learning on the contextual features of these variables.

We present data collected from participants on Amazon Mechanical Turk who annotated kitchens with locations of items. We describe several methods for enabling a learning robot to predict item–location ratings for novel users from this data. We also discuss how these ratings can be used in an interactive system that attempts to find a trade-off between asking the user for the correct input and making a wrong choice. We demonstrate results on a collection of user-annotated, simulated kitchen examples.

2. RELATED WORK

Researchers have long argued that user interfaces can benefit from learning from their users [10]. However, devices in the home like the Nest thermostat [11] require a significant amount of learning effort. Yang et al. [21] reported that early users struggled with understanding what the Nest thermostat learned and that these users found it hard to override its learned behavior.

Home organization is a good example of a task that both exhibits strong user preferences and is highly desirable to automate. Several researchers found that cleaning and organization are the top two tasks users most want robots to do [20, 4]. Pantofaru et al. [13] argue that, unlike cleaning tasks, organization is "nuanced and emotional." They also found that, even if they could afford the help, many people will not hire human assistants to organize their belongings because it is too personal of a problem. Therefore, a robot learning this task must tread carefully by weighing the cost of requiring too many user interactions and misplacing items.

Several researchers in the robotics community have begun to look at placing and organizing items. Cha et al. [5] examined methods for predicting locations of items in a user's kitchen from other items in the kitchen. Using item-related features like item-type and use, they found that a Support Vector Machine classifier (SVM) performed the best in their domain.

However, in collaborative filtering, SVM methods are often disregarded due to their poor performance on sparse datasets. We present as a baseline a Support Vector Regression method that struggles on the full data set when considering individual locations, but performs better only over item and location features. However, our proposed solutions still outperform this SVR method.

Fisher et al. [6] presented a probabilistic model that can generate plausible scenes of items from user-provided examples. Leveraging a larger scene database, they can create arrangements of novel items suitable for the user. Collaborative filtering methods extend beyond this type of generative model by incorporating preferences of other users to predict given user's preferences. Jiang et al. [7] discuss a method for placing items optimizing for stable and semantically relevant locations, but do not capture a user's preferences for locations among semantically identical locations. Schuster et al. [19] learn organizational principles to place items into meaningful locations, but do not make use of user preferences to select locations. However, a robot utilizing our algorithms could also easily incorporate these capabilities of identifying free space and stable placement poses within the specific location chosen by our prediction system.

The work by Abdo et al. [2] most closely mirrors the contributions of our paper. They use collaborative filtering techniques on item–item pair ratings and then approximately solve a minimum k-cut problem to best group the items from their predicted ratings. Using these groupings, they then place the items into semantically identical bins or shelves. We build on this work to solve the problem of predicting specific locations for items in a kitchen, as opposed to just their general groupings. We present several methods to solve this more general problem, including a method to adapt their technique over collocation data to predict suitable locations. We further present a context-aware collaborative filtering technique that leverages item–item collocation information without its explicit representation in the training data.

3. METHOD

Our proposed solutions predict a user's preferred locations for items they want organized. It consists of two components, a rating prediction collaborative filter and an algorithm for producing the optimal location given a set of location ratings produced from the collaborative filter.

There are two solutions we present here. The first builds on previous work by using a collaborative filter to predict a user's item–item ratings – whether the two items should be placed together or not. Predicting the correct location is a matter of choosing the location which contains the item with lowest distance rating to the item being placed for the evaluated user. The second uses a context-aware collaborative filter to directly predict a user's item–location ratings. The predicted location is just the location with the minimum item–location rating for the user.

We envision this work as a component in a robot's back and forth interactions with their users. To minimize the number of interactions required at the risk of misplacing objects, we also present a method for tuning the predictive success of the location-prediction system by allowing the system to choose to place based on its predicted value or ask the user for the actual location.

4. RECOMMENDER SYSTEMS

Recommender systems are concerned with the problem of predicting user–item ratings. That is, given user u, what rating $r \in \mathbb{R}$ would they assign to the item i? For the problem of household organization, we modify the problem so that the system makes predictions on item–item ratings or item–location ratings. This modification makes our problem a variant of the classical recommender problem in that users provide ratings on either item–item pairs or item–location pairs.

Collaborative filtering is a category of methods for recommender systems that seek to predict ratings for items novel to a user from ratings of the items from similar users. The typical input is a rating matrix $R \in \mathbb{R}^{M \times N}$, where rows are associated with the items, and the columns are associated with users. Factorization methods attempt to generate a lower dimensional representation to improve generalization.

As shown in Abdo et al. [2], we decompose this matrix into:

$$R = B + \bar{R} \tag{1}$$

where B is a bias matrix that encodes the global bias, item bias and user bias, and $\bar{R}$ is the residuals matrix that the collaborative filtering algorithm attempts to learn. We use the L-BFGS minimization algorithm [9] to find the factorization of the residual matrix and bias matrix that minimizes the regularized squared error.

4.1 Context-aware recommender systems

Compared to classical recommender systems, context-aware recommender systems assume that some additional information exists that relates to the user's choices of ratings. For example, a user may choose a location based on the other items already placed there or based on the location's salient features. We include information about the locations into the context-based recommender, but our chosen context-aware recommender also learns automatically information related to item–item collocation to inform its predictions.

4.2 Factorization Machines

Rendle et al. [15] developed factorization machines to make predictions in a model with all multi-degree interactions among the context variables with sparse data in linear time. For a system that models only interactions between two variables at a time, the model equation is:

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\mathbf{v_i} \cdot \mathbf{v_j})\, x_i x_j, \tag{2}$$

where $\mathbf{x}$ is the input variable vector, w_0 is the global bias, $\mathbf{w}$ is the weights over $\mathbf{x}$. $(\mathbf{v_i} \cdot \mathbf{v_j})$ is the dot product of $\mathbf{v_i}$ and $\mathbf{v_j}$ and models the interaction of the i-th and j-th variable, where $\mathbf{v_{i/j}} \in \mathbf{V}$ is a factorized representation of the weighting of $x_{i/j}$.

Instead of estimating an individual weight parameter ($w_{i,j} \in \mathbb{R}$) for higher-order interactions ($d \geq 2$), the factorization of $\mathbf{v}$ enables modeling even under sparsity. Here, w_0, $\mathbf{w}$ and $\mathbf{V} \in \mathbb{R}^{\mathbf{n} \times \mathbf{k}}$ are the parameters to be estimated.

Context does not need to be explicitly represented because it can be learned through the factorization of $\mathbf{v_i}$. Though we show that the item–item collocation has very powerful explanatory power, we do not need to represent that information in the context variables. Instead, we include users, items, locations and features, and the optimization of the model will seek to explain these ratings appropriately.

Though the model presented in equation 2 suggests a solution would require a run time of $\mathcal{O}(kn^2)$ where k is the number of dimensions and n is the number of context variables, they show this equation can be reformulated into a linear solution with runtime complexity $\mathcal{O}(kn)$. Indeed, for models with higher degrees of interactions, it can still be refactored into a linear time solution.

5. LOCATION SELECTION

One difficulty of choosing a correct location for an item is that many locations in a kitchen are interchangeable and are justifiable placement locations ignoring the current organization. A user's preferred organization therefore requires the robot to identify the most likely location from existing item placements and physical suitability of the location for the item.

For our work, we chose to code ratings as 0.0 being a positive example and 1.0 as a negative example, though the reverse – 0.0 for negative and 1.0 for positive – is equally valid. For item–location pairs, a rating, $r(i, l) = 0.0$ indicated the item is in the preferred location and 1.0 indicated it is not. For item–item pairs, a rating, $r(i, i') = 0.0$, indicated the items should be located together. Rating predictions produced by the collaborative filtering methods are real valued approximately in the range of $[0, 1]$.

5.1 Choosing a Location from Item–Item Pairs

In our first approach, item–item ratings $r(i, i')$ are learned from training input via a context-unaware collaborative filtering technique [2]. Finding a suitable location for an item in the kitchen then requires finding the location that results in the best match to the item–item ratings. If a location is empty, the method takes the mean pair rating as its default.

Formally, for items $i, i', j, j' \in I = \{i_1, i_2, \cdots, i_m\}$, and location $l \in L = \{l_1, l_2, \cdots, l_n\}$ where $i \neq i'$ and $j \neq j'$, n is the number of locations and m is the number of items, we want to find the item–location rating $R_{\text{item–item}}$ summarized from item–item ratings $r(i, i')$.

$$R_{\text{item–item}}(i, l) = \begin{cases} \min_{i' \in l} r(i, i') & \text{if } \exists i' \in l, \\ \frac{1}{m^2} \sum_{j \in I} \sum_{j' \in I} r(j, j') & \text{otherwise.} \end{cases} \tag{3}$$

We then select the location with the lowest distance rating. We denote the event of item i being placed in location l as $i \in l$ and write the selected location as

$$\hat{l} = \underset{l}{\operatorname{argmin}}\, R_{\text{item–item}}(i, l). \tag{4}$$

5.2 Predicting Location Ratings through FMs

We use the factorization machine's model to predict item–location ratings. Each row in the input matrix includes context variables for the user, item, location, and a location category. For each provided example of an item–location pairing in the input data, we assign a distance rating of 0. We produce $n - 1$ other ratings of value 1 for this same item in all the other locations in the kitchen. Each location variable is encoded uniquely for the user, even if in experiments different users annotated the same kitchens.

In Equation 5, we show the context variable vector we used in our dataset. Each context variable is assigned a value of 1 when active and 0 otherwise. In our notation, u is a user, i_i is an item, l_j is location, and $\Phi(l)$ are the feature values over the active location variable l:

$$\mathbf{x} = \{u_1, \cdots, u_U, i_1, \cdots, i_m, l_1, \cdots, l_n, \Phi(l)\}. \tag{5}$$

Because the model directly learns item–location ratings $R_{\text{item–location}}(i, l) = r(i, l)$, we can find the location with the lowest distance rating:

$$\hat{l} = \underset{l}{\operatorname{argmin}}\, R_{\text{item–location}}(i, l). \tag{6}$$

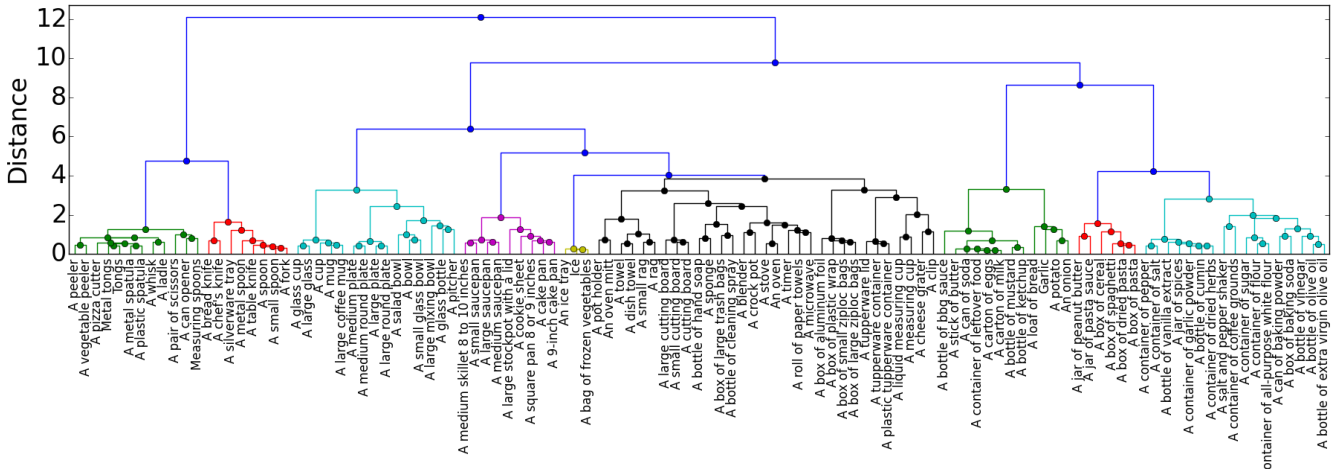


Figure 1: A hierarchical clustering of 111 kitchen items and their mean pair ratings. Groups were colored for distances less than or equal to 4

5.3 Combination Method

Both techniques, predicting via item–item pairs and item–location pairs, provide ratings for an item in a location. We can combine the two values in a number of ways. We chose harmonic mean to capture information provided by both ratings without being unduly influenced by the relative offset biases between the two types of ratings.

$$\hat{l} = \underset{l}{\operatorname{argmin}} \left(\frac{2.0 R_{\text{item–location}}(i,l) * R_{\text{item–item}}(i,l)}{R_{\text{item–location}}(i,l) + R_{\text{item–item}}(i,l)} \right). \quad (7)$$

6. DATA COLLECTION

To create a data set with maximum opportunities for generalizing between users, we were interested in finding a collection of items that many people would be familiar with and would have in their households.

6.1 Finding common kitchen items

To create our item set, we began by pulling a list of recommended cooking items from 'How to Cook Everything' by Bittman [3]. Their recommended items consisted of cooking tools and basic foods useful for cooking a variety of recipes. We also examined the list of items identified by Cha et al. [5] in their CMU kitchen dataset. They surveyed several households and annotated all the items found in their participants' kitchens. We merged the items across these sources and removed duplicates—items that we judged to be the same as an item already in the list—and were left with 398 items. We replaced items commonly found in a container, like seasonings, oils or other liquids, with the appropriate container of that item ('olive oil' became 'a bottle of olive oil').

We randomly split the selected items into surveys of 10 items, and one survey of 8 item and asked participants on Amazon Mechanical Turk if they had any of the items in their household kitchen. For each survey, we collected 20 responses. Participants were allowed to answer as many different surveys as they chose (min: 1, max: 30, mean: 10.5). No one participant completed all the surveys.

From the full set, we selected the items at least 85% of respondents indicated were in their households, producing a list of 111 items to be used for location annotations.

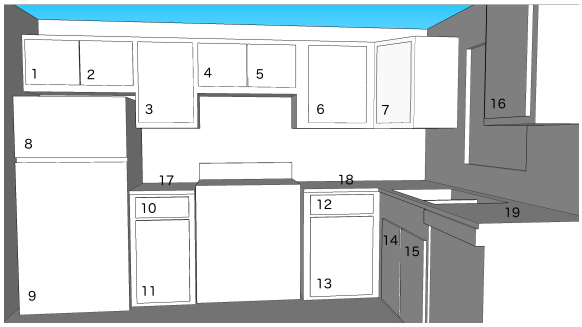


Figure 3: Diagram of a kitchen with numeric labels.

6.2 Kitchen annotations

We were provided four kitchen layouts from the authors of the CMU Kitchen Dataset. We labeled each drawer, cabinet, refrigerator, freezer and counter with a unique numerical identifier. Figure 6.2 shows one of the labeled kitchen images we provided for a survey. We created four different surveys, each with a different labeled kitchen. We asked participants to assign locations in their labeled kitchen for all 111 items. Participants were asked not to respond to more than one survey. For the participants who took multiple surveys, we only kept responses to their first. We received 25 responses to each survey. We removed three responses from two groups due to duplication of users. Ten participants from the previous task also completed this one.

To summarize patterns in how objects were placed closely to one another, we built the hierarchical clustering of item–item pair distances shown in Figure 1. Over all users, and for each item–item pair, we computed the fraction of times the items were not placed together. Groups of items with distances less than or equal to 4 are given unique colors in the diagram. The figure illustrates that items fall into clear categories in their placement around the kitchen. For example, 'a glass cup', 'a large glass', 'a cup', 'a mug', and 'a large coffee mug' were commonly placed together.

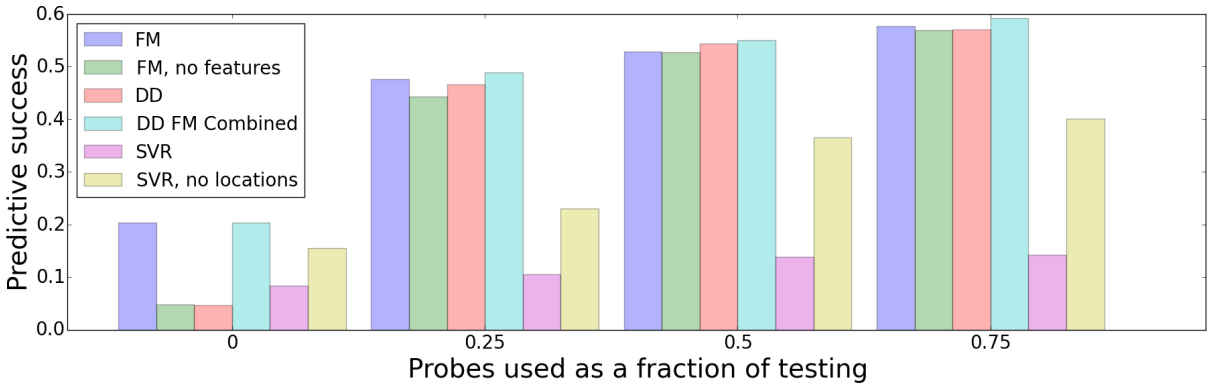


Figure 2: Prediction accuracy versus the number of probes for several different learning methods.

7. EXPERIMENTS

From the data presented in Section 6.2, we created four sets of training and testing data by assigning data collected for three kitchens to training and the fourth to testing. Even though many people annotated the same kitchen, we used a variable for a location unique to each user. For each item and location in a kitchen, we created a collocation data set by indicating for each item–location pair in the original data set whether it was found with one of the other 111 items. That is, we changed the data lines from rating–item–location to rating–item–item–location. We set the associated rating to 0.0 if the two items were found together in the location and 1.0 otherwise.

For the item–location dataset, probes were drawn from the testing data. For each user in the testing data, we selected a random set of items and moved the ratings of that user for those items to the training dataset. We used the same sets of items for the collocation dataset and created a set of probes for all item pairings in this set of items.

7.1 Model configuration

We trained the factorization-machine model (FM) using libFM [16] on the training data to produce predictions for each user–item–location in the testing dataset. We used adaptive stochastic gradient descent with a learning rate of 0.03, a sampling standard deviation of 0.1 and 30 for the number of dimensions of the model. The item–location tuple with the lowest predicted rating was chosen as the predicted location of the item.

For the data-driven factorization method (DD), we trained the collaborative filter on the collocation training dataset without the additional probes. The probes were then used to update the model through the new-user update method presented by Abdo et al. [2]. As done in their work, we also built our model with three dimensions for the factorization. The limited memory BFGS minimization algorithm from SciPy [8] was used to find the optimal variables, with a stopping criteria 'factr' set to 10.

We included a support vector regression (SVR) model to predict ratings from the dataset. The SVR model was trained on the item–location data set. We used the implementation provided in scikit-learn [14]. The user columns of the item–location dataset were removed to reduce sparseness. We used a radial basis function kernel, with the default gamma of $\frac{1}{N}$ where N is the number of features. The stopping tolerance was set to 0.0001, and no shrinkage was used.

We trained both SVR models and one FM model with features. For both, we included features about the locations describing the location type: drawer, low cabinet, high cabinet, counter, refrigerator, freezer. Additionally, for the SVR models, we also included item features related to the use of the item: edible, drinkable, food storage, etc.

For the combined model, we used the FM model trained with features and the DD method discussed above.

7.2 Results

Figure 2 presents the predictive success of the non-probe items in the testing dataset for two FM models, a data-driven factorization model over item collocation data (DD), a model combining the FM model and DD model (DD FM Combined), and two SVR models. We measured predictive success as the fraction of times the model predicted the correct location a user chose to place an item.

For the two FM models, we present a model trained as we described in previous sections (FM), and also a model trained on these variables but without location features (FM, no features). For the two SVR methods, we show the results of one model trained using location variables (SVR) and one without (SVR, no location). We show results for different fractions of items used as probes, specifically 0, $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$, corresponding to 0, 28, 56, and 83 probes out of the 111 total items.

As seen in this graph, the proposed methods perform quite strongly for this difficult task. Even with no information about a current user (0 probes), the FM and DD FM combined model can correctly predict a significant fraction of locations for that user. The FM, no features and the DD methods perform strongly when probes are available, but reduce to random guessing when 0 probes are used. This is due to the limitations of collaborative filtering without additional contextual information. Without any information from the user, the base performance is generally quite poor.

The FM models both performed strongly over a wide range of probes. Both of the FM models were trained without explicit collocation data, yet both performed as well as the data-driven model, implying that the FM model is correctly learning the importance of item–item pairings for prediction. Indeed, the success of the DD method to predict locations when probes were available illustrates the importance of the inherent item–item affinities presented earlier.

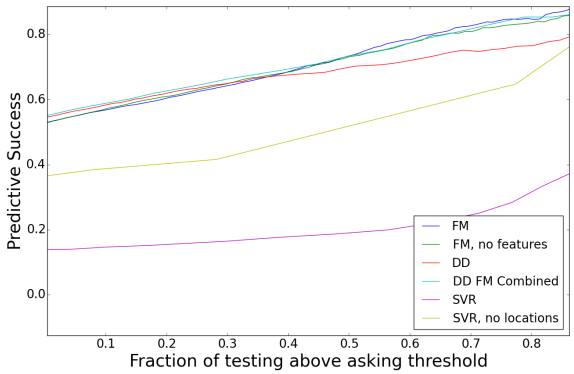


Figure 4: Predictive success versus the fraction of testing that did not meet the minimum required threshold for placement. Lower fractions imply more locations were chosen, with less placement confidence.

The DD FM Combined model outperformed the FM models and the DD model by themselves when probes are provided. It combined the learning of the FM model on location features with the high accuracy of item–item ratings provided by the DD model.

Both SVR methods performed worse than the proposed solutions regardless of the number of probes presented. The SVR method trained with locations performed especially poorly. The locations were uniquely encoded for each individual user, which created a highly sparse input that posed difficulty to this method. By excluding locations, the SVR no locations method performed better because this input is significantly less sparse. Regardless, the collaborative filtering methods strongly outperformed the SVR methods.

7.3 Placement-versus-asking trade-off

An important design consideration for an interactive system is the trade-off between placing an item incorrectly and asking the user for the correct placement. We observed that each algorithm places higher confident answers closer to the ranges of their output. Therefore, we can enable systems to ask questions for uncertain items by requiring a threshold for ratings.

Figure 4 plots the success of each algorithm against the fraction of items in the testing dataset that do not meet the threshold for placement with 50% of probes already placed. We normalized the predictions of each algorithm between 0 and 1 to simplify the comparison.

Data points at a fraction of 0 correspond to data presented in Figure 2 for 50% probes. The DD FM Combined method remains among the strongest overall. Data for fractions above 0.85 were not included because there are too few data points for accurate predictive success.

8. CONCLUSIONS

In this work, we presented several methods for reducing the burden of personalization that comes from users supplying a learning system household organization preferences. We collected data about the existence, pairing and location of items in participants' kitchens. We found that people have a lot of items in common but also many unique items not found in other kitchens. A system that recognizes a location suitable for an item could easily still fail to match the user's preference for its location. In response, we presented a solution that incorporates a user's preferences but enables an autonomous system to make use of perceivable features of these locations when learning.

We showed that a context-based collaborative filtering model called factorization machines is well suited to predict item–location ratings. Choosing the appropriate location is a simple task of finding the location with the best rating. For interactive systems, we show the value in trading off asking for a location versus placing it incorrectly.

Though household robotics entails a sizable burden of personalization, this issue also plagues many other modern smart-household devices like thermostats and lighting. Machine-learning methods that use data from other households in learning a user's preferences will require smart recontextualizations that can map the preferences drawn from a collection of different users onto the user's own household space. We envision these topics as bright areas of future research.

References

[1] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard. Collaborative filtering for predicting user preferences for organizing objects. *arXiv preprint arXiv:1512.06362*, 2015.

[2] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard. Robot, organize my shelves! tidying up objects by predicting user preferences. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1557–1564. IEEE, 2015.

[3] M. Bittman. *How to Cook Everything: 2,000 Simple Recipes for Great Food*. John Wiley & Sons, 2011.

[4] M. Cakmak and L. Takayama. Towards a comprehensive chore list for domestic robots. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 93–94. IEEE Press, 2013.

[5] E. Cha, J. Forlizzi, and S. S. Srinivasa. Robots in the home: Qualitative and quantitative insights into kitchen organization. In *HRI*, pages 319–326, 2015.

[6] M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics (TOG)*, 31(6):135, 2012.

[7] Y. Jiang, C. Zheng, M. Lim, and A. Saxena. Learning to place new objects. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3088–3095. IEEE, 2012.

[8] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed May 19, 2016].

[9] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[10] P. Maes et al. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.

[11] Nest. Home | nest. https://nest.com/, 2016. Accessed: 2016-4-1.

[12] U. Panniello, A. Tuzhilin, M. Gorgoglione, C. Palmisano, and A. Pedone. Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 265–268. ACM, 2009.

[13] C. Pantofaru, L. Takayama, T. Foote, and B. Soto. Exploring the role of robots in home organization. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 327–334. ACM, 2012.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] S. Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[16] S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.

[17] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 635–644. ACM, 2011.

[18] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[19] M. J. Schuster, D. Jain, M. Tenorth, and M. Beetz. Learning organizational principles in human environments. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3867–3874. IEEE, 2012.

[20] C.-A. Smarr, T. L. Mitzner, J. M. Beer, A. Prakash, T. L. Chen, C. C. Kemp, and W. A. Rogers. Domestic robots for older adults: attitudes, preferences, and potential. *International journal of social robotics*, 6(2):229–247, 2014.

[21] R. Yang and M. W. Newman. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 93–102. ACM, 2013.

A Cross-Cultural Analysis of Explanations for Product Reviews

John O'Donovan
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
jod@cs.ucsb.edu

Shinsuke Nakajima
Faculty of Computer Science and Engineering
Kyoto Sangyo University, Kyoto, Japan
nakajima@cse.kyoto-su.ac.jp

Tobias Höllerer
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
holl@cs.ucsb.edu

Mayumi Ueda
[3]Faculty of Economics
University of Marketing and Distribution Sciences, Kobe, Japan
Mayumi_Ueda@red.umds.ac.jp

Yuuki Matsunami
Faculty of Computer Science and Engineering,
Kyoto Sangyo University,
g1245108@cc.kyoto-su.ac.jp

Byungkyu Kang
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
bkang@cs.ucsb.edu

ABSTRACT

Cosmetic products are inherently personal. Many people rely on product reviews when choosing to purchase cosmetics. However, reviewers can have tastes that vary based on personal, demographic or cultural background. Prior work has discussed methods for generating attribute-based explanations for item ratings on cosmetic products, based on associated text-based reviews. This paper focuses on evaluating explanation interfaces for product reviews and related attributes. We present the results of a cross-cultural user study that evaluates five associated explanation interfaces for cosmetic product reviews across groups of participants from three different cultural backgrounds. We applied a 3 by 2 within subjects experimental design in a user study (N=150) to evaluate effects of UI design and personalization on a range of user experience metrics in a cosmetics shopping scenario. Results of the study show that 1) Korean and Japanese speakers chose the most complex UI more often than English speakers. 2) older participants also preferred more options in cosmetic product selection, regardless of cultural background. 3) personalization of product ratings did not show an effect on user experience. 4) Attribute-based explanations were preferred over star-ratings for all three cultures. 5) Rating propensity evaluation showed that Japanese provided significantly higher ratings than Korean or English participants, and that Females provided higher ratings than Males, regardless of background.

CCS Concepts

•**Human-centered computing** → *HCI design and evaluation methods; User models; User studies;*

Keywords

User Experience, Explanation, Decision Making, User-Centric Evaluation

1 Introduction

Over the last 25 years, recommender systems have attempted to help users find the right information at the right time [15]. More recently, the proliferation of e-commerce applications supports buying and selling products in the global market with relatively little effort. Increasingly, consumers are relying on customer reviews to inform purchasing decisions [8]. In many cases, product reviews are presented in summary form via mechanisms such as star ratings. Such representations, however, typically fail to capture the subtle opinions that exist in the accompanying text-based reviews. In this paper, we build on recent work that automatically extracts attributes and associated ratings from online product reviews [10]. In particular, we focus on understanding how visual representations of various types of extracted item ratings impact user experience and conversion likelihoods in an e-commerce setting, as exemplified in Figure 1. Motivated by recent research that shows the importance of user experience over traditional accuracy metrics in recommender systems [7], we conduct a user experiment to understand how rating display affects user experience. Specifically, we applied a 3 by 2 within subjects design (Table 1) in an online study (N=150) to evaluate effects of UI design and personalization on a user experience metrics in a cosmetics shopping scenario, considering the following research questions:

R1: Do cross-cultural preference differences exist for recommendation interfaces? If so, what are the key predictors of these differences?

R2: Are there cross-cultural preference differences for personalized v/s non-personalized recommender system interfaces?

R3: Are there cross-cultural preference differences between traditional (star-rating) and more granular attribute-based recommender system interfaces?

R4: Are there differences in rating propensities across the

three cultures? If so, what are the strongest predictors of observed rating shifts?

The cosmetics domain was used for this study, since they are sold globally and are inherently personal in nature. To explore variances in opinions on the explanation interfaces across different cultural backgrounds, participant groups were sourced from American, Japanese and Korean cultural backgrounds. These particular groups were selected as a representative sample with diverse cultures, and because they are among the fastest growing markets for cosmetics.[1]

2 Related Work

In this study, we focus on explanations and transparency of recommender systems and on the (associated) role of product attributes mined from product reviews. Here, we discuss several related work in these areas.

Product Attributes To understand consumer behavior in economics, research has focused on the different attributes and uncertainties that consumers consider when purchasing a product [8, 13]. For buyers, these attributes play important roles when deciding to purchase a product. More importantly, attributes vary widely across product types and users' personal tastes. For example, [3] study the effects of search attributes and provide a comparison between traditional and online supermarkets. A recent study on description and performance uncertainty [4] focused on the difficulty in assessing the product's characteristics. Building on works such as [13] that show advantages of using fine-grained product attributes in the recommendation process, we aim to further our understanding of the role of fine-grained product attribute ratings in consumer decisions.

Explanation and Transparency in Recommendation Within the recommender systems research community, there is an increasing understanding of the need for user-centered evaluations [12]. Recent keynote talks [2] and workshops [14] have helped to highlight the importance of this topic. In this paper, we follow Knijnenburg et al.'s [9] argument for a framework that takes a user-centric approach to recommender system evaluation, beyond the scope of recommendation accuracy. In contrast to that work however, we argue that decision quality is an important evaluation metric that goes beyond the user experience metrics described in [9], and further, that it can be used to explain observed usage patterns for search and recommendation tools. Garcia-Molena [6] described differences and similarities between search and recommendation, and argued that interactive interfaces can help users understand and use these tools in more efficient ways. Along the same vein, it has also been recognized that many recommender systems function as *black boxes*, providing no transparency into the working of the recommendation process, nor offering any additional information to accompany the recommendations beyond the recommendations themselves [7]. To address this issue, static or interactive/conversational explanations can be given to improve the transparency and control of recommender systems. Research on textual explanations in recommender systems to date has been evaluated in wide range of domains (varying from movies to financial advice [5]). From a cross-cultural perspective, Pu and Chen performed a related study that evaluated perceptions of different recommendation interfaces in [1], using subjects from Chinese and Swiss backgrounds. In contrast to their study, which compared a novel UI against a list view and assessed user experience metrics, we focus on the perception of attribute ratings versus traditional less-fine grained ratings, and on the impact of personalization on these perceptions. A second contrast to [1] is that our work explores rating propensity across the different groups.

[1]http://polishcosmetics.pl/Korean-Market-Analysis.pdf

3 Mining Attribute Ratings

This study builds on a recent work [10] on attribute extraction from online product reviews. Specifically, we posit that more explanations of a given product in the form of multiple attributes with corresponding scores (on five star rating scales), see Figure 1, can provide benefits to potential customers. In the prototype of the proposed recommender system, both personalized information (*Simgroup* ratings: "Users similar to you rate this item as") and multiple product attributes extracted from a review text are added as features. Through an online user study, we apply both novel approaches as controlled variables to the prototype design and investigate the preference of the users to such features across demographic backgrounds, particularly, cultural backgrounds (English, Korean and Japanese).

4 Interface Design

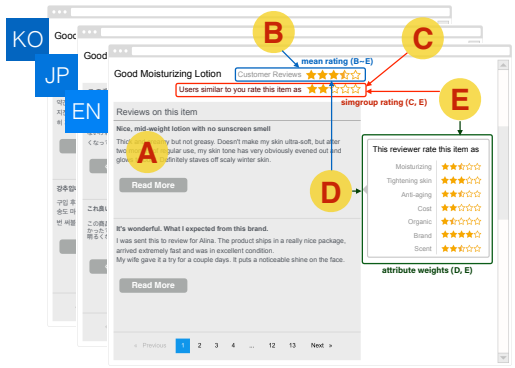


Figure 1: Screenshots of the interface used in the online user study. The annotations A-E show the items that varied in each condition, as shown in Table 1.

We designed a novel user interface for product review pages based on the feedback we received from a preliminary user study (N=100). We performed the study with a simple design layout to test the different visual conditions outlined in Table 1. Participants gave feedback on their preference for each UI in a virtual shopping scenario. They were also required to leave a comment on the interface design. For example, they reported the benefit of the new features, such as "*I like the level of detail it has related to the product*", and suggested preferred features, such as "*More alive colors*" / "*More explanations and ratings*". The collection of 100 comments were manually assessed, and improvements were made to the UI, including shortened review text with "read more" button and breakdown of multiple attributes extracted from the review text on stars. The revised design is shown in Figure 1.

5 Experimental Setup

Figure 1 shows an example of the refactored interface for a sample product review. To test our hypotheses above, a

Table 1: Overview of the controlled variables for the online user study.

UI Config	non-personalized (no information from similar users)	personalized (with social data from similar users)
review text only	**A**: product review text	
review text with star rating	**B**: A + mean rating on stars	**C**: A + mean rating and the rating from active user's *simgroup* on stars
review text, star rating and attributes	**D**: B + attribute weights computed from current review text (on stars)	**E**: C + attribute weights computed from current review text (on stars)

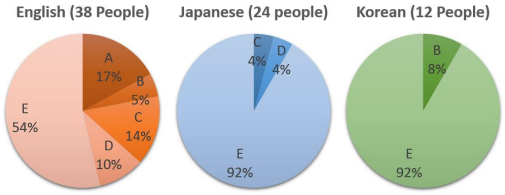


Figure 2: Preferred User Interface by Culture.

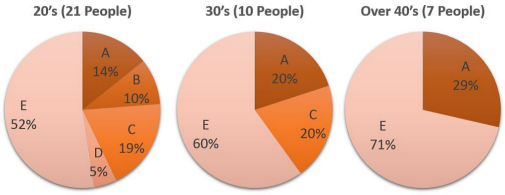


Figure 3: Preferred User Interface by Age.

3x2 within subjects experiment was conducted, controlling for personalization, and rating type, as shown in Table 1. The study (N=150) was performed on the crowdsourcing platform, Amazon Mechanical Turk. Each participant was shown a randomly ordered set of 5 different design layouts corresponding to the treatments in Table 1, and were asked to rank them in order of preference. They were also asked to rate the helpfulness of each. Participants were evenly balanced across cultural backgrounds. All participants were shown with the five interfaces in random order. The content was shown in their primary language based on their cultural background. Overall, participants took between 5-10 minutes doing the study, and were paid $1.50 for their time. Questions were added to test for user attention level and for language proficiency, including identification of differences between UIs and simple math questions written in the appropriate language. After filtering our data based on these metrics, group sizes were 39, 25 and 12 for English, Japanese and Korean, respectively. Participant age ranged between 18-64 with an average of 26. Gender groups were not evenly distributed, as expected for the cosmetics domain, with 70% female and 30% male.

6 Results

Perception and Rating Differences Figure 2 shows the results for the UI ranking task, broken down by age. The result shows a clear preference for design E in all groups, but there is a significant increase in that preference for participants over 40 (shown on the right side). This effect was also seen from 100 participants in the preliminary study. Interface E, shown in Figure 1, shows the most information, and allows users to understand how users similar to them rate individual product attributes. This effect might be a result of specific preferences for cosmetics developing with age, and

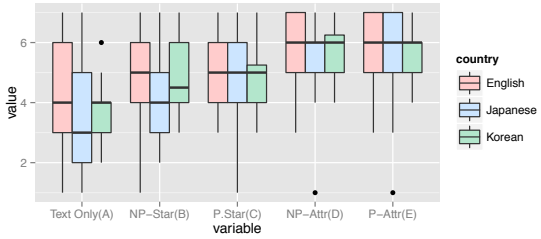


Figure 4: Cross-cultural perspective of helpfulness of the five evaluated interfaces.

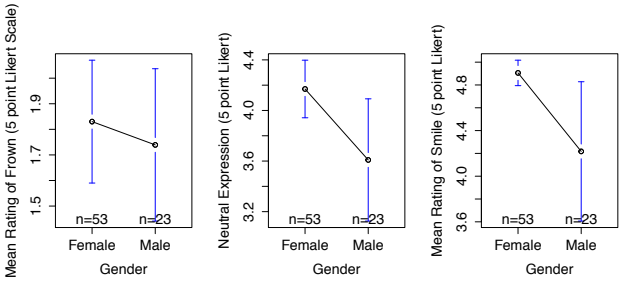


Figure 5: Difference in rating propensity by gender.

accordingly, an increased need to explore user ratings on fine grained product attributes (see Figure 3).

Personalization and Rating Type Figure 4 shows the results of perceived usefulness of the interfaces, broken down by cultural groupings. Each UI condition is shown as a group on the x-axis, and each group contains the mean utility score for the three cultural groups. The x-axis groups (UI treatments) are also ranked from left to right based on number of visible features (UI complexity). This graph shows several interesting effects: first, there is a general preference across all groups for the attribute-based representations (groups D and E, on the right side), over less granular, star-ratings or text-based UIs. This is a promising result that indicates that attribute extraction and visualization has a positive effect on Ux. The second interesting result is that within the star-rating group (2nd and 3rd group) and the attribute-rating (4th and 5th) groups there is no notable difference between the personalized and non-personalized treatments. This result tells us that the granularity of presented ratings has more positive impact on user experience than the perception that the ratings come from similar users. To investigate this result in more depth, a followup experiment is planned with a large corpus of product reviews collected from Amazon.com [11] [2] to compute actual similarity scores based on user profiles. This would clearly give better insight into the observed effect. Figure 4 also answers R2, in that there are no significant differences between the cultural groups within

[2]http://jmcauley.ucsd.edu/data/amazon/

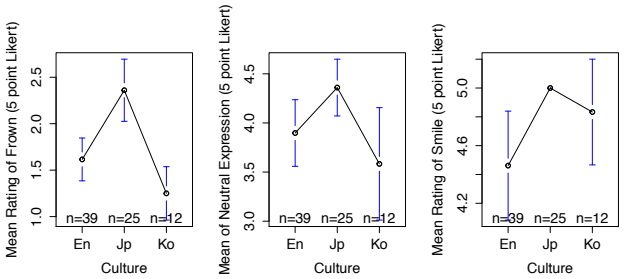


Figure 6: Difference in rating propensity by culture.

each UI treatment, although the Japanese showed a trend towards favoring the more complex UI treatments.

Rating Propensity For some users, rating an item with a specific number of stars can have very different meanings. User ratings on items serve as the basis for most collaborative recommendation techniques, but they tend to ignore such differences when computing neighborhoods for recommendation. Further, little work has been done to understand cross-cultural differences in rating propensity. Since these participant groupings were available our experimental setup, a logical step was to evaluate rating propensities within each of the cultural groups, to serve as both an independent result, and as a weighting factor for the analysis in Figure 4. Each participant was shown three randomly ordered faces, showing expressions with happy, neutral and sad expressions. They were asked to rate the 'happiness' perceived in each on a five point Likert scale. Figure 5 shows the results by gender (for all groups). Interestingly, there is a trend for Females to rate higher than males, and the difference becomes more pronounced for the 'happy' expression, shown on the rightmost plot of Figure 5 with a mean difference of 0.7 (relative increase of 16%, $p<0.005$). Figure 6 shows the results of the rating propensity analysis broken down by cultural group. Again, the graphs represent mean rating for sad, neutral and happy expression ratings from left to right, respectively. Here, we see a clear trend for higher ratings in the Japanese group across all three expressions. While this is only a small-scale initial study, we believe that this is an important result for the study of recommender system performance across different cultures in general, and a follow-up study on propensity of ratings for recommender systems is planned to investigate this further.

7 Discussion and Future Work

This study applied a 3 by 2 within subjects experimental design in a user study (N=150) to evaluate effects of UI design and personalization on a range of user experience metrics in a cosmetics shopping scenario using participant groups from three different cultural backgrounds. Results of the study show that 1) Korean and Japanese speakers chose the most complex UI more often than English speakers. 2) older participants also preferred more options in cosmetic product selection, regardless of cultural background. 3) personalization of product ratings did not show an effect on user experience. 4) attribute-based explanations were preferred over star-ratings for all three cultures. 5) Rating propensity evaluation showed that Japanese had significantly higher ratings than Korean or English, and that Females provided higher ratings than Males, regardless of background. A clear next-step is to evaluate on real product data. The authors plan a follow-up study to compare LDA and dictionary-based approaches to product attribute extraction, and to explore how the resulting attributes can improve explanations, and user profiles for collaborative filtering. Additionally, a more detailed evaluation of the different rating propensities across cultures is underway using a larger number of participants and multiple product domains.

8 References

[1] L. Chen and P. Pu. A cross-cultural user evaluation of product recommender interfaces. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 75–82, New York, NY, USA, 2008. ACM.

[2] E. H. Chi. Blurring of the boundary between interactive search and recommendation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 2–2. ACM, 2015.

[3] A. M. Degeratu, A. Rangaswamy, and J. Wu. Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of research in Marketing*, 17(1):55–78, 2000.

[4] A. Dimoka, Y. Hong, and P. A. Pavlou. On product uncertainty in online markets: Theory and evidence. *Mis Quarterly*, 36, 2012.

[5] A. Felfernig, E. Teppan, and B. Gula. Knowledge-based recommender technologies for marketing and sales. *Int. J. Patt. Recogn. Artif. Intell.*, 21:333–355, 2007.

[6] H. Garcia-Molina. Thoughts on the future of recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 1–2. ACM, 2014.

[7] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[8] Y. Kim and R. Krishnan. On product-level uncertainty and online purchase behavior: An empirical analysis. *Management Science*, 61(10):2449–2467, 2015.

[9] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[10] Y. Matsunami, M. Ueda, S. Nakajima, T. Hashikami, S. Iwasaki, J. O'Donovan, and B. Kang. A method for automatic scoring of various aspects of cosmetic item review texts based on evaluation expression dictionary. In *Proceedings of the 24th International MultiConference of Engineers and Computer Scientists*, IMECS '16, pages 392–397. IAENG, 2016.

[11] J. McAuley and A. Yang. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. *ArXiv e-prints*, Dec. 2015.

[12] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, 2006.

[13] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod. Extracting and visualizing trust relationships from online auction feedback comments. In *IJCAI*, pages 2826–2831, 2007.

[14] J. O'Donovan, N. Tintarev, A. Felfernig, P. Brusilovsky, G. Semeraro, and P. Lops. Joint workshop on interfaces and human decision making for recommender systems (intrs). In H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, editors, *RecSys*, pages 347–348. ACM, 2015.

[15] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, pages 175–186, 1994.