Exploring users' perception of rating summary statistics

Coba, Ludovik; Zanker, Markus; Rook, Laurens; Symeonidis, Panagiotis

DOI
10.1145/3209219.3209256

Publication date
2018

Document Version
Final published version

Published in
Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization

Exploring Users' Perception of Rating Summary Statistics

Extended Abstract

Ludovik Coba
Free University of Bozen
Bolzano, Italy
lucoba@unibz.it

Markus Zanker
Free University of Bozen
Bolzano, Italy
mzanker@unibz.it

Laurens Rook
TU Delft
Delft, The Netherlands
L.Rook@tudelft.nl

Panagiotis Symeonidis
Free University of Bozen
Bolzano, Italy
psymeonidis@unibz.it

ABSTRACT

Collaborative filtering systems heavily depend on user feedback expressed in product ratings to select and rank items to recommend. These summary statistics of rating values carry two important descriptors about the assessed items, namely the total number of ratings and the mean rating value. In this study we explore how these two signals influence the decisions of online users based on choice-based conjoint experiments. Results show that users are more inclined to follow the mean indicator as opposed to the total number of ratings. Empirical results can serve as an input to developing algorithms that foster items with a, consequently, higher probability of choice based on their rating summarizations or their *explainability* due to these ratings when ranking recommendations.

KEYWORDS

Recommender systems, User studies, Explanation styles

ACM Reference Format:
Ludovik Coba, Markus Zanker, Laurens Rook, and Panagiotis Symeonidis. 2018. Exploring Users' Perception of Rating Summary Statistics: Extended Abstract. In *UMAP '18: 26th Conference on User Modeling, Adaptation and Personalization, July 8–11, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3209219.3209256

1 INTRODUCTION

User ratings are one of the key ingredient to collaborative filtering algorithms to automatically assess how likely items might match users' tastes. Although, recently, implicit signals on users' actual behavior have turned out to possess even more predictive power for practical systems [4, 6], ratings still play a dominant role in constructing the value and quality perception of an item in the eyes of online consumers [2]. Collaborative explanations [3] provide justifications for recommendations by displaying information about the rating behavior of a users' neighbourhood, as has been already identified by Herlocker et al. [5]. Also, e-commerce sites usually provide at least rating summary statistics along with the products in their catalogs.

This extended abstract therefore discusses a study that explores how the two dominant characteristics of a rating summarization, namely the number of ratings and their mean value, impact the choice behavior of users. Results show that - all things being equal - users are clearly biased towards selecting items with higher means as opposed to larger numbers of ratings, which provides clear indications about the degree of *persuasiveness* [12] of collaborative explanations for different products and different user neighborhoods. Note, that a full-length paper including a full description of the methodology and all results can be accessed in [1].

UMAP '18, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5589-6/18/07.
https://doi.org/10.1145/3209219.3209256

2 RELATED WORK

Explanations for recommendations have received considerable research attention over the past years [3, 11]. There are different ways of explaining recommendations based on collaborative filtering mechanisms as presented in Herlocker et al. [5]. They explored 21 different interfaces and demonstrated that specifically the "user" style improves the acceptance of recommendations. The "user" style of explanation provides information about the neighborhood, which is determined based on a generic notion of similarity between users when analyzing their observed behavior or expressed opinions (i.e., buys, clicks, ratings etc.).

In this work we are interested in shedding light on users' trade-off between rating numbers and their mean values when they have to make a choice.

Conjoint analysis is a market technique suitable for revealing user preferences and trade-offs in the decision making process[9]. Conjoint analysis has successfully been employed in a wide range of areas, such as education, health, tourism, and human computer interaction.In the field of recommender systems and online decision support, Zanker and Schoberegger [13] employed a ranking-based conjoint experiment to understand the persuasive power of different explanation styles over the users' preferences.

To the best of our knowledge, the persuasive effect of the characteristics in rating summarizations has not yet been studied. The conjoint methodology as employed in market research for decades represents a best practice in order to quantify the perceived utility of the characteristics of different rating summarizations.

3 METHODOLOGY AND DESIGN

We perform an experimental user-study in order to understand the trade-off mechanisms between confrontation with different configurations of rating summarizations. We base our analysis on the Choice-Based Conjoint (CBC) methodology, which is also denoted as Discrete Choice Experiments by several authors [8]. In conjoint designs, products (a.k.a., *profiles*) are modeled by sets of categorical or quantitative *attributes*, which can have different *levels*. In CBC experiments, participants have to repeatedly select one profile from different *sets of choices*, which nicely matches real-world settings when users are confronted with recommendation lists.

Figure 1: An example snapshot of a choice set, with three different rating summary profiles based on different attribute levels.

Table 1: Probability of choice over profiles in decreasing order.

	# of Ratings	Mean Rating	Pr. of choice	Utility
1	Large	High	35.47 %	3.31
2	Medium	High	23.73 %	2.90
3	Small	High	20.80 %	2.77
4	Large	Average	6.65 %	1.63
5	Medium	Average	4.45 %	1.23
6	Low	Average	3.90 %	1.10
7	Large	Low	2.22 %	0.53
8	Medium	Low	1.48 %	0.13
9	Small	Low	1.30 %	0.00

We used a 3 x 3 choice experiment, where 3 different levels of mean values and of number of ratings have been defined in order to build 9 different summary statistics. Formally, a rating summary statistic is a frequency distribution on the class of discrete rating values. We choose the movie domain for our study and employed the Netflix dataset[4] to identify representative real-world levels for characterizing rating frequency distributions. In addition, variance and skewness of the rating frequency distributions is controlled for, by fixing them with the median values from the respective Netflix rank distributions, which are 1 for variance and -0.5 for skewness.

Our CBC design consisted of $N = 6$ choice sets with $m = 3$ alternatives (see Figure 1). The design was generated and evaluated using SAS MktEx macros [7]. The SAS code for replicating and evaluating the survey is accessible for download[1].

Between January and February 2018 a group of *54 people* were invited to participate in our choice experiments. The participants were presented with the following hypothetical situation:

"Assume that you find yourself in the situation that you need to make a choice between three movies to watch on a movie platform. These three movies are equally preferable to you with respect to all other movie information you have access to (title, plot, actors etc.). Other users' ratings are aggregated and summarized by their number of ratings, the mean rating value and their distribution. Therefore, we would like to know your choice, by solely considering these ratings summary statistics."

4 RESULTS

Detailed results and an extensive discussion is provided in [1]. There was a clear and statistically significant preference relation over the three levels for mean rating values. However, in terms of the total number of ratings, users did not seem to care that much.

From the different levels of preference weights (partial utilities) for our two signals (i.e. levels of the profile attributes) we can also derive the perceived overall utility (see Table 1). The probability of selecting any of the 9 profiles was computed and ordered by decreasing values in Table 1. Changes in mean value were well and strongly perceived, while the number of ratings had far less impact on users' choice - i.e., an increase in the mean rating value by one level increased the probability of choice by a factor of three to four, when everything else was kept constant.

[1]SAS code: https://github.com/ludovikcoba/CBC;

5 DISCUSSION

Rating summarizations provide important clues to users in online choice situations. Marketing research has shown that consumers are strongly guided by online reviews, and that the mean rating value is interpreted as an indicator for the quality of a product [2]. Also in our study, participants seem to have been following this quality hypothesis.

The total number of ratings, on the other hand, is typically regarded as an indicator for the popularity of a product or an item in general. Given that with larger sample sizes, all things being equal, the mean rating value becomes more informative, it is also very reasonable that, in case of a large number of ratings, users would be more likely to follow this choice. This work is in line with prior research on the effects of potential decision biases such as position, decoy or framing effects, on the choice behavior of users [10] and it can be purposefully exploited to develop more persuasive systems [12].

REFERENCES

[1] Ludovik Coba, Markus Zanker, Laurens Rook, and Panagiotis Symeonidis. 2018. Exploring Users' Perception of Collaborative Explanation Styles. (5 2018). http://arxiv.org/abs/1805.00977

[2] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. Do online reviews matter?-An empirical investigation of panel data. *Decision support systems* 45, 4 (2008), 1007–1016.

[3] Gerhard Friedrich and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32, 3 (2011), 90. https://doi.org/10.1609/aimag.v32i3.2365

[4] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.

[5] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*. 241–250. https://doi.org/10.1145/358916.358995

[6] Dietmar Jannach, Lukas Lerche, and Markus Zanker. 2018. *Recommending Based on Implicit Feedback*. Springer International Publishing, Cham, 510–569. https://doi.org/10.1007/978-3-319-90092-6_14

[7] Warren Kuhfeld. 2005. Experimental design, efficiency, coding, and choice designs. *Marketing research methods in sas: Experimental design, choice, conjoint, and graphical techniques* (2005), 47–97. https://support.sas.com/techsup/technote/mr2010c.pdf

[8] Jordan J Louviere, Terry N Flynn, and Richard T Carson. 2010. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* 3, 3 (2010), 57–72.

[9] Vithala R Rao. 2014. Choice Based Conjoint Studies: Design and Analysis. In *Applied Conjoint Analysis*. 127–183. https://doi.org/10.1007/978-3-540-87753-0{_}4

[10] Erich Christian Teppan and Markus Zanker. 2015. Decision Biases in Recommender Systems. *Journal of Internet Commerce* 14, 2 (2015), 255–275.

[11] Nava Tintarev and Judith Masthof. 2015. Explaining recommendations: design and evaluation. In *Recommender Systems Handbook*. Springer US, Boston, MA, 217–253. https://doi.org/10.1007/978-1-4899-7637-6

[12] Kyung-Hyan Yoo, Ulrike Gretzel, and Markus Zanker. 2012. *Persuasive recommender systems: conceptual background and implications*. Springer Science & Business Media.

[13] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *CEUR Workshop Proceedings*, Vol. 1253. 33–36. http://ceur-ws.org/Vol-1253/paper6.pdf