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

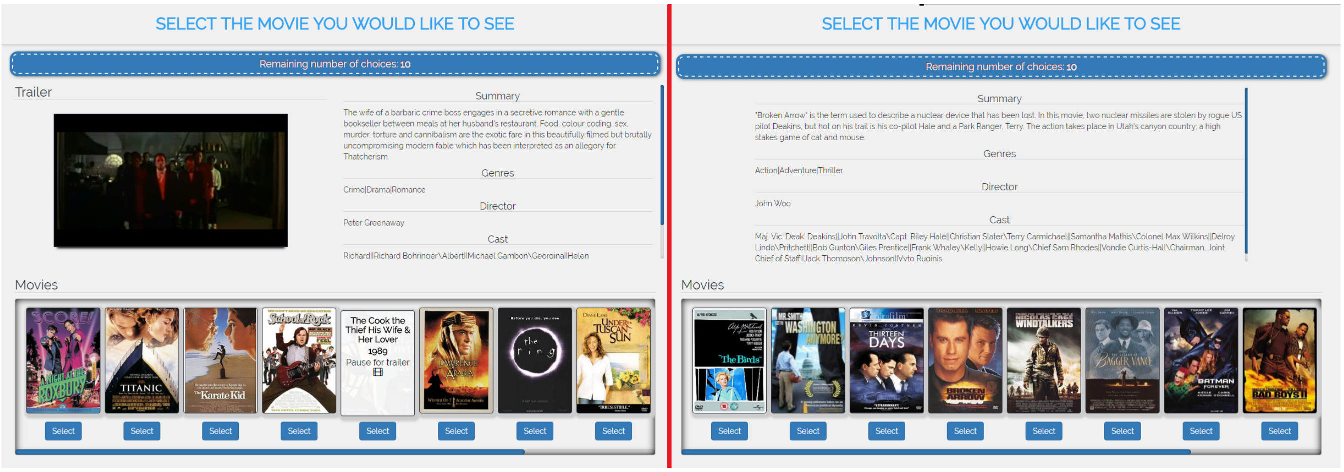

Figure 1: User Interface used for the study. The interface to the left is used for participants in the trailer condition, the interface to the right for participants in the non-trailer condition. Within the interface the list of items to choose from is shown below, the trailer and additional metadata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

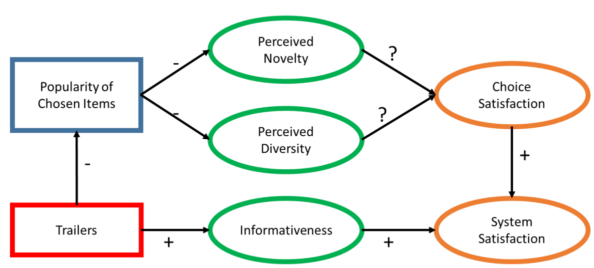

Figure 2: Path diagram of expected effects.

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.

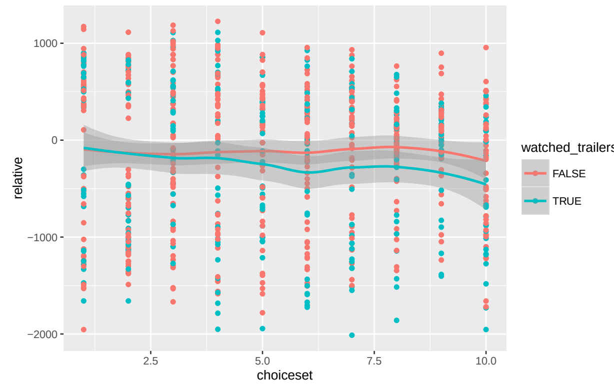

Figure 3: Relative Rank of Chose Item as a Function of Choice Set, for participants that watched trailers (Green) and those that did not (Red).

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

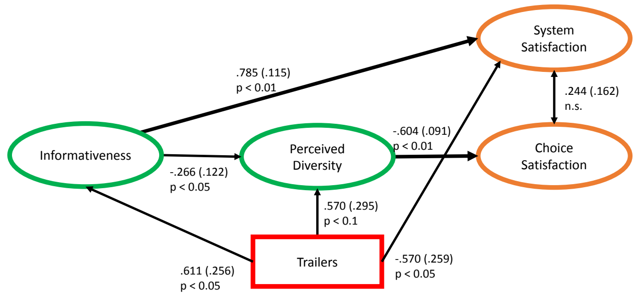

Figure 4: Path model of the CFA. Width of the arrows show effect sizes, numbers next to the arrows show the standardized effect size, with standard error and significance levels.

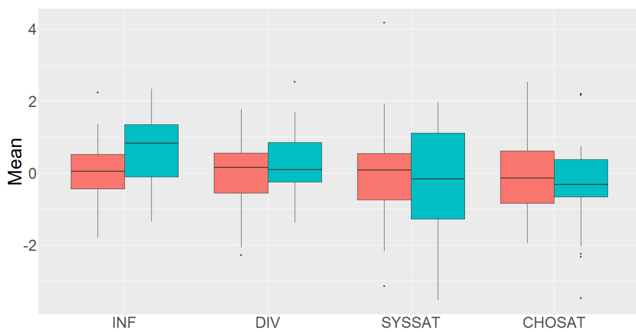

Figure 5: Boxplots of the estimated marginal means for the perceived informativeness (INF), perceived diversity (DIV), system satisfaction (SYSSAT) and choice satisfaction (CHOSAT), for participants in the trailer (red bars) and non-trailer (green) conditions.

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.