Quality and Diversity of Recommender Systems Users' Choices: a Simulation Perspective

Discussion Paper

Naieme Hazrati, Francesco Ricci

Free University of Bolzano, Bolzano, Italy

Abstract

Recommender Systems (RSs) generate personalised suggestions for items and can influence collective users' choices behaviour. The impact of operational RSs on users' decisions can be assessed by analysing the actual choices' diversity, and quality, e.g., the users' satisfaction for their choices. But, in order to estimate the potential impact of an RS in new scenarios and for not yet deployed RSs, simulating user-system interactions can be valuable. We here illustrate a simulation framework consisting of users, items, and alternative RSs. We simulate users' choices over consecutive time intervals, by assuming that an RS influences the users' choices with recommendations. We measure global properties of the simulated choices, such as their diversity and quality. The obtained results, and the proposed simulation framework, can be used by a system designer in order to anticipate the effect of a candidate RS in its long term usage.

Keywords

recommender system, simulation, choice behaviour, diversity

1. Introduction

Recommender Systems (RSs) literature has shown that these information filtering techniques can profoundly affect individuals' choices [1]. Hence, nowadays, there is a growing attention and concern to better understand how RSs can bias collective users' choice behaviour. This important topic has been studied mostly by relying on off-line analysis of their performance and, more recently, by developing algorithmic simulations of the users' choices for items when users are also exposed to recommendations. Important metrics, such as the diversity and the choice quality of the bulk of simulated choices have been considered [2, 3, 4, 5, 6]. These studies have obtained some interesting results showing their validity and importance for understanding RSs' effect on users [2, 3]. However, they are far from being complete in the analysis of the effect of alternative usage conditions of the RS, e.g., the number of recommendations or the types of RSs. Moreover, their reliability in properly simulating realistic usage settings could be improved. For instance, some previous simulation studies made critical modelling simplifications of the user-recommendation interactions, such as, assuming that the users can only choose recommended items, or that users have simple preference models, which are not correctly estimated from the observation of their past behaviour [3, 2].

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

nhazrati@unibz.it (N. Hazrati); fricci@unibz.it (F. Ricci)

In this short contribution, we illustrate a flexible simulation framework, initially proposed in [4], that copes with some of the identified limitations: simple and unrealistic set of possible choices; use of synthetic data sets of users and items; not taking into account the dynamic of new users and new items. The simulation framework described in this paper addresses these issues by leveraging existing data sets of real logged users' choices (Amazon data sets), in order to properly analyse the effect of an RS. We simulate an iterative choice-making procedure of a collection of users in the presence of recommendations produced by alternative RSs, as well as, when no RS is used. Simulated users make choices for a time interval (a month), and then the RS is retrained by also considering the data of the simulated choices. This simulation is repeated for a certain number of consecutive time intervals, while new items and new users can also enter the system. The user's simulated choices are influenced by the users' utility for the items, which are estimated by a de-biased rating prediction model [7]. The user's simulated choices are among the items in her 'awareness set', which contains the items that she is supposed to know, plus the recommended items, which are added to the user's awareness set. The persuasive effect of the RS is simulated by increasing the perceived utility of the recommended items, which makes the recommendations more likely to be chosen.

Using the above-mentioned simulation framework, we focus on the following important and yet not clearly answered research questions:

- ***RQ1***: How personalised and non-personalised RSs affect the evolution of choice diversity? What features of the RSs determine their specific impact?

- ***RQ2***: Do personalised RSs suggest items that users rate higher than non-personalised RSs?

- ***RQ3***: Does a larger users' awareness of the catalogue of the items, i.e., being aware of a larger number of items, lead to better choices, that is, higher users' rating for the choices?

In the remaining part of this article we first outline the most important modeling aspects of the proposed simulation framework and we then illustrate its usage, by leveraging behaviour data stored in two Amazon data sets.

2. Simulation

We simulate the iterative process of users' choice making for items. Users select items in monthly time intervals. We use true users' choices data, collected by an eCommerce platform (Amazon), up to a certain time point t_0, as starting point of the simulation. We use this initial data set to train an RS, and then we simulate the users' choices in successive months. At the end of each month, the RS is re-trained by adding to the training data also the simulated choices of that month. To simulate users' choices, we estimate the users' preferences for the items (ratings). This is performed with another model, different from the RS, which computes an unbiased prediction of the ratings that the simulated users would give to the items [7]. This prediction model is trained by using the full data set of true users' choices, in order to model as correctly as possible the preferences of the real users, which are here simulated. This 'unbiased' prediction model disentangles the predicted ratings from observation biases, leading to predicted ratings that better represent users' intrinsic preferences.

In each month interval, the users select items one after another. When a user u is simulated to make a choice, first her awareness set A_u is built. This is the set of items that the user is supposed to know and that can choose. We assume that A_u contains a fixed number of items (e.g., 2000) among the most popular ones (most frequently chosen previously) and with the largest estimated ratings for u. In other words, we assume that u has some knowledge of the items' catalogue, which is not derived from the recommendations; this knowledge is assumed to be influenced by available information on the items and by the user's preferences. Then, an RS suggests a set of items to u (50 in our study), which are added to the user's awareness set. Finally, the user makes a choice based on a multinomial logit choice model (MLM). We adopt this model because it is a simple but effective approach, which has been previously validated. In MLM a user's choice is drawn by using a probability distribution over the possible choices (i.e., the items in A_u):

$$p(u\ chooses\ i) = \frac{e^{v_{ui}}}{\sum_{j \in A_u} e^{v_{uj}}}$$

Here, v_{ui} is the utility of the item i, and, if the item is not recommended, v_{ui} is proportional to the estimated rating of the item $\hat{r}_{ui}$. Conversely, if the item i is recommended, the utility is supposed to be larger, i.e., $v_{ui} \propto \delta * \hat{r}_{ui}$, with $\delta > 1$. That is, if an item is recommended, the persuasive effect of the RS is modelled as if it is giving to the simulated user the impression that the item is more valuable, and therefore, according to MLM definition, it is also more likely to be chosen.

Five rather different RSs are studied in our simulation:

- *PCF* - Popularity-based CF: is a nearest neighbourhood collaborative filtering RS that suggests the most popular items among the choices of nearest neighbour users [2].
- *LPCF* - Low Popularity-based CF: is similar to *PCF*, but it penalises the score of popular items, computed by *PCF*, by multiplying it with the inverse of their popularity [2].
- *FM* - Factor Model: is a collaborative filtering RS based on matrix factorization [8].
- *POP* - Popularity-based: is a non-personalised RS that recommends the most popular items to the users.
- *AR* - Average Rating: is another non-personalised RS that recommends items with the highest average ratings.

We analyse the simulated choices by considering metrics that are computed on the set of all the simulated choices of the users, and make clear the variety of the effects of RSs on users' choice behaviour: (a) the *Gini* index of the chosen items [2]; (b) *Choice Coverage* of the catalogue (percentage) [3]; (c) *Popularity* of the chosen items; (d) Average predicted rating of the choices (*Choice's Rating*) [3]; (e) *Recommendation Coverage* of the catalogue (percentage); (f) Probability of acceptance of the recommendations (*Recommendation Acceptance*).

We use two *Amazon* collection's data sets to conduct the simulation: *APPS* and *Games*. [9]. These data sets contain timestamped ratings of users for items distributed on several months. Ratings signal actual choices, as are normally provided after the purchase. In our analysis, we simulate choices performed in the last ten months (of the recorded data), while using the previous months' data to bootstrap the simulation with the required observations of previous users' choices.

3. Results

A selection of the simulation results are shown in Table 1, where the values of the considered metrics, are computed at the end of the ten months of simulated choices. LPCF produces a high diversity (lower Gini), comparable to when no RS is used (Apps data set) and even larger than that (Games data set). Hence, in practice, only relying on methods that explicitly penalise popular items, as in LPCF, an RS can increase choice diversity compared with a baseline where no RS is influencing users' choices. In general, personalised RSs produce a larger choices diversity than non-personalised ones, which instead encourage choices over more popular items (Popularity metric) and cover a smaller part of the catalogue (Choice Coverage metric).

Table 1
Diversity and quality of the simulated choices in Amazon *APPS* and *Games* data sets.

	Gini						Choice Coverage						Recommendation Coverage				
RS →	**PCF**	**LPCF**	**FM**	**POP**	**AR**	**No RS**	**PCF**	**LPCF**	**FM**	**POP**	**AR**	**No RS**	**PCF**	**LPCF**	**FM**	**POP**	**AR**
Apps	0.82	0.76	0.80	0.89	0.91	0.74	0.50	0.56	0.52	0.42	0.38	0.60	4.28	0.43	0.62	0.03	0.01
Games	0.89	0.83	0.89	0.96	0.96	0.89	0.20	0.27	0.20	0.13	0.13	0.191	4.43	0.24	0.41	0.04	0.01
	Choice's Ratings						Popularity						Recommendation Acceptance				
RS →	**PCF**	**LPCF**	**FM**	**POP**	**AR**	**No RS**	**PCF**	**LPCF**	**FM**	**POP**	**AR**	**No RS**	**PCF**	**LPCF**	**FM**	**POP**	**AR**
Apps	4.06	4.00	3.98	4.18	4.28	3.97	0.00035	0.00021	0.00024	0.00216	0.00232	0.0001	**0.41**	**0.34**	**0.34**	**0.49**	**0.56**
Games	4.25	4.18	4.18	4.39	4.43	4.25	0.00028	0.00018	0.00021	0.00132	0.00138	0.0002	0.51	0.41	0.43	0.57	0.57

In the second research question, we ask if personalised RSs produce better choices, i.e., with higher Choice's Rating, compared to non-personalised RSs. Surprisingly, non-personalised RSs result in choices for items with a larger predicted rating than personalised RSs. Hence, non-personalised RSs can be strong baselines if the goal is to nudge users to choose items that they will like. This means that if there is no need to diversify the choices, a non-personalised RS may suffice.

In the third research question, we ask if having a larger awareness set size, i.e., a better knowledge of the items' catalogue, results in a higher Choice's Rating. The results, not presented here for lack of space, show that when awareness set grows, choices are more diverse, and there is also a clear decrease in the acceptance of the recommendations, which leads to choices with smaller Choice's Rating. Hence, being aware of more items does not help users to make better choices but helps them to make more diverse ones. This result is due to the fact that if the users make choices among the items belonging to a larger set of options then there is an increased probability to choose among less good items, but more diverse.

4. Conclusion

We have illustrated a simulation framework that is able to produce a simulated, but realistic, succession of users' choices for items. Users' preference data are extracted from a data set of observed choices. Users are supposed to make choices among two types of items: that they are likely to know and that an RS explicitly suggest to them. Users' simulated choices are determined by a choice model that is based on the estimated utility of the items. This approach enables to study the effect of RSs on users choices going beyond previous studies that were limited to the analysis of properties and biases of the recommendation algorithms alone [10]. We have obtained interesting findings, such as, non-personalised RSs can produce choices that are rated higher than those produced by personalised RSs. Moreover, we found that choices on average are rated lower when the awareness set size of the simulated users is larger.

These, and other findings not here described for lack of space, shed light on the complex effect of exposing users to recommendations. The practical value of this study is related to possibility to anticipate, without deploying an RS, the potential effect of the RS on the users' choices. Hence, by using the proposed framework, developers can better pre-select candidate RS algorithms, hence also reducing potential undesired negative effects and the system.

References

[1] F. Ricci, L. Rokach, B. Shapira, Recommender systems: introduction and challenges, in: Recommender systems handbook, Springer, 2015, pp. 1–34.

[2] D. Fleder, K. Hosanagar, Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity, Management science 55 (2009) 697–712.

[3] Z. Szlávik, W. Kowalczyk, M. Schut, Diversity measurement of recommender systems under different user choice models, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[4] N. Hazrati, M. Elahi, F. Ricci, Simulating the impact of recommender systems on the evolution of collective users' choices, in: Proceedings of the 31st ACM Conference on Hypertext and Social Media, 2020, pp. 207–212.

[5] S. Yao, Y. Halpern, N. Thain, X. Wang, K. Lee, F. Prost, E. H. Chi, J. Chen, A. Beutel, Measuring recommender system effects with simulated users, arXiv preprint arXiv:2101.04526 (2021).

[6] D. Bountouridis, J. Harambam, M. Makhortykh, M. Marrero, N. Tintarev, C. Hauff, Siren: A simulation framework for understanding the effects of recommender systems in online news environments, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 150–159.

[7] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, Recommendations as treatments: Debiasing learning and evaluation, in: international conference on machine learning, PMLR, 2016, pp. 1670–1679.

[8] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: 2008 Eighth IEEE International Conference on Data Mining, Ieee, 2008, pp. 263–272.

[9] R. He, J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: proceedings of the 25th international conference on world wide web, 2016, pp. 507–517.

[10] J. Huang, H. Oosterhuis, M. de Rijke, H. van Hoof, Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems, in: Fourteenth ACM Conference on Recommender Systems, 2020, pp. 190–199.