

Simulation-Based Evaluation of Interactive Recommender Systems

Behnam Rahdari¹, Peter Brusilovsky¹

¹*School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA*

Abstract

Offline data-driven evaluation is considered a low-cost and more accessible alternative for the online empirical method of assessing the quality of recommender systems. Despite their popularity and effectiveness, most data-driven approaches are unsuitable for evaluating interactive recommender systems. In this paper, we attempt to address this issue by simulating the user interactions with the system as a part of the evaluation process. Particularly, we demonstrate that simulated users find their desired item more efficiently when recommendations presented as a list of carousels compared to a simple ranked list.

1. Introduction

For many years, empirical evaluation based on various kinds of user studies was the key approach for evaluating all kinds of user-adaptive systems, i.e., interactive systems that can adapt their behavior to individual users [1]. While user studies could be considered as an ultimate way to assess and compare any user-centered systems, these studies are known as very expensive. It is also a challenge to obtain user study data on a sufficient scale to reliably distinguish specific user modeling and personalization approaches. In response to these challenges, several research fields that could be considered as sub-areas of user-adaptive and personalized systems established *data-driven* approaches for evaluating systems in these areas. For example, data-driven evaluation of learner modeling in personalized education systems is based on large collections of student problem-solving traces. The ability to better predict a learner's success in these traces is considered a sign of better-quality modeling [2, 3]. Similarly, data-driven evaluation of recommender systems is based on the large volume of user past rating data. The ability to better approximate user rating or position positively rated items higher in the ranked list is considered a sign of better-quality recommendation [4, 5].

The establishment of data-driven evaluation approaches was very important for the recommender system field. Promoted by the Netflix prize, these approaches helped to engage a large number of researchers in the work on recommender systems and stimulated rapid progress in the development and evaluation of recommendation algorithms. Data-driven evaluation quickly became a gold standard in the field overshadowing the empirical evaluation approaches.

InTRS'22: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 18-23, 2022, Seattle, WA

✉ ber58@pitt.edu (B. Rahdari); peterb@pitt.edu (P. Brusilovsky)

🌐 <https://sites.pitt.edu/~ber58/> (B. Rahdari); <https://sites.pitt.edu/~peterb/> (P. Brusilovsky)

🆔 0000-0001-6514-912X (B. Rahdari); 0000-0002-1902-1464 (P. Brusilovsky)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Numerous papers discussed the comparative benefits of data-driven vs. empirical evaluation and pointed out that these studies frequently deliver different results [5, 6, 7]. The proponents of data-driven evaluation stressed the opportunity to obtain large-scale data and to evaluate new ideas relatively fast, especially given the increasing number of available datasets. The proponents of user studies stress that the end-user is the ultimate judge and that the ability to assess many “beyond precision” aspects of recommendation is not possible without engaging users. It is currently accepted that data-driven evaluation is not a replacement for empirical evaluation, rather the two approaches, frequently referred to as off-line and online evaluation, are complementary and together could offer a more complete picture in assessing and comparing recommender systems [8, 5]. In other words, it is important to have a choice between data-driven evaluation and user studies when evaluating recommender systems.

However, the choice between data-driven and empirical evaluation approaches is currently not available for researchers working on various *interactive recommender systems* [9], which present recommendation results in a more complex way than a ranked list and engage users in different forms of interactions. The key problem here is that user behavior in these systems is more complex than in recommender systems based on a ranked list. User work with traditional ranked lists is well-explored and user tendency to examine the list from the top and favor top-ranked items is well known [10, 11]. These observations helped to create commonly accepted metrics for offline evaluation such as nDCG [12] or MRR [13]. Traditional metrics, however, are not applicable to interactive recommender systems. These systems might have multiple ranked lists or no ranked list at all and their effectiveness is defined by the whole user interaction rather than a single output of recommendation results. Does it mean that offline data-driven evaluation is not an option for interactive recommenders?

This paper attempts to make a case for simulation-based evaluation of an interactive recommender system. The idea of this approach is a continuous simulation of user behavior in a target system while computing various performance metrics “on the go”. It could be applied to relatively complex interaction scenarios as long as user behavior in these scenarios could be modeled sufficiently well. While this approach enables the application of diverse metrics that are typically used in empirical studies, it is based on simulated rather than real users and can be performed offline. Simulation-based evaluation is a recognized approach for evaluating various kinds of interactive systems [14, 15], however, its application for evaluating recommender systems is still an exception [16, 17].

To demonstrate the power of the simulation-based approach and its potential value for data-driven evaluation of interactive recommender systems, the paper presents a simulation-based study that compares user behavior and performance in two types of recommendations interfaces - an interface with multiple carousels (sometimes referred to as a *multilist*) and a traditional ranked list. This simulation-based study was enabled by the availability of user behavior models for both, the ranked list and the multilist interfaces. While the main goal of this paper was to demonstrate how a simulation-based offline study could be organized in a recommender system context, we also pay attention to the study results, which help to explain the increasing popularity of carousel-based interfaces.

The paper is structured as follows. In Section 2, we review past research on simulation-based evaluation. In Section 3, we present an empirically grounded model of user behavior in ranked list-based and carousel-based interfaces. In Section 4, we present in detail the setup for the

simulation-based study. Finally, we discuss our results in Section 5.

2. Related Work

As mentioned in the introduction, simulation-based evaluations have been used in a number of fields where sufficiently detailed models of user behavior could be built. Simulation-based evaluation is a recognized approach for evaluating various kinds of personalized interactive systems from adaptive learning systems [15] to personalized information access systems [14, 18]. The goals of simulation-based evaluation differ between application areas and frequently depend on the reliability of behavior models that support the simulation. On one end of the spectrum are cognitively grounded behavior models that are supported by studies of human cognition and confirmed by empirical studies. A well-known example is SNIF-ACT model [14] that simulates user behavior in hypertext navigation. This model is based on Information Scent theory [19] and was used to assess the quality and navigability of Web sites without real users. Popular “artificial student” models [15] used for evaluation of adaptive educational systems also belong to this group. On the other end, there are a range of simple behavior models [20] that might not be able to reliably predict the details of user behavior but could be useful to explore a range of “what if” scenarios in assessing the impact of various interface augmentations.

Early attempts to use simulations for exploring information filtering and recommender systems were made in the first decade of 2000 [18, 21], however, it took another 10 years for this approach to become truly noticed and used in this field [16, 22]. While the role of simulation-based research in the recommender system context is currently recognized, simulations are most frequently used for the exploration of recommender systems rather than their evaluations. The most popular research direction enabled by simulation is examining the impact of a recommender system, as a whole, on various aspects of user behavior [21, 23, 17, 20]. This work is typically enabled by the user choice models [22]. While research on click models reviewed in more detail in Section 3 offers a solid ground for simulation-based studies, there were few cases where models of user click behavior are used for comparative offline evaluation of recommender system design options. A notable exception is the work of Dzyabura and Tuzhilin [16] who used simulation to compare an interface based on a combination of search and recommendation to interfaces based on search or recommendation alone. However, this work used a relatively simple behavior model that was not based on empirical observations or theory. In our work, we would like to specifically focus on the opportunities that simulation-based evaluation offers to advance research on interactive recommender systems while emphasizing the need for reliable empirically grounded behavior models.

3. Carousels Versus Ranked List: The Models

The main obstacle in using a simulation-based approach for evaluating interactive recommender systems is the need for sufficiently reliable behavior models for the realistic simulation of user behavior. While these models do not yet exist for all kinds of interactive recommender systems, user behavior in several types of interactive recommender systems is explored sufficiently well to build these models. In our paper, we want to demonstrate the use of simulation-based evaluation

for assessing the efficiency of carousel-based recommendation interfaces and comparing it with the traditional recommender list approach.

The goal of our simulation-based study is to compare user performance with two types of recommendations interfaces - a carousel-based multi-list and a traditional ranked list - in a typical modern recommendation context where items could be associated with multiple “interests” and users could favor several of these interests in parallel (although probably to a different extent and at a different time). Depending on the domain, these interests could have different semantic natures. For example, it could be movie genres (such as *action movies*) or research topics (such as *context-aware recommendation*). For uniformity, we refer to these interests as *topics*. Note that some recommender systems could model interests as latent categories rather than explicit semantic topics. In this paper, we focus on domains with explicitly represented interests, to separate the problem of latent interest discovery from the problems of user modeling and item ranking.

3.1. Ranked List Interaction Model

Extensive studies of user information access behavior started in the field of information retrieval and were originally motivated by the need to improve Web search engine performance. While “old school” information retrieval considered item relevance as the only factor determining user decision to click on a specific result, it became increasingly evident that the position of items in a ranked list has to be considered as well [24]. A sequence of eye-tracking studies with users of search engines [10, 25, 11, 26, 27] helped understand how users explore a ranked list of results, recognize the impact of item positions and build a range of so-called *click models* [28, 29, 30, 31, 32]. These click models attempted to explain the user behavior by a generative model, which can be learned from data. The most popular of these models known as *the cascade model* [29, 30] assumes that the user examines the list of recommended items from top to bottom until they find an attractive item. After that, they click on that item and leave satisfied. This seemingly simple model explains the position bias in recommender systems, that lower-ranked items are less likely to be clicked than higher-ranked items. In turn, this information can be used to de-bias logged data [33], or to learn better ranking policies either offline [34] or online [35, 36].

We define our ranked list interaction model on the basis of the *cascade model*. This means that we assume the user starts by examining the first item on the list and continue the examination one by one until finding the desired item or when there is no item left to examine. However, to make this interaction model consistent with the carousel interaction model, in our simulations we adapted the ranked list interaction model to a 2D context. We define the ranked list as a matrix of $m \times n$ recommended items, which is examined row by row (Figure 1a). This isolates the impact of 2D presentation from the impact of topic labels.

The ranked list interaction model in a 2D context defines as follows: The user starts at position (1, 1). If that item is not desired, the user proceeds to the next item (1, 2). The user examines row 1, from left to right, until the desired item is found or the end of the row is reached. If the end of the row is reached, the user moves to the item (2, 1), the first item in the next row. Then the user examines this row, from left to right, and this process continues until the desired item is found.

3.2. Carousel Interaction Model

A recommender interface with multiple carousels offers its users several ranked lists, each marked with a topic, in place of a single ranked list. This interface leaves the choice of the current topic of interest to the user making the recommendation process more interactive. While this interface is not very complex, evaluating it in a traditional offline way using static metrics for its [37]. As a result, carousel-based interfaces are predominantly explored through user studies [38, 39]. On the other hand, the studies of multi-list and other 2D presentation interfaces [38, 39, 40] provided useful information for developing models of user behavior in carousel-based interfaces.

The key work on examining user behavior in 2D presentation interfaces was performed not in the area of recommender systems, but in the area of Web search as an extension of the work on click models. As the presentation of results was becoming increasingly two-dimensional, understanding user behavior when exploring 2D presentation was important to optimize novel ways to present information in 2D [41, 42, 40]. A range of user studies that frequently engaged eye-tracking brought consistent results. There was compelling evidence that the users examine the 2D presentation top-down, row-by-row, similarly to the case of a 1D ranked list. In the horizontal dimension, each row is examined as far as to ascertain whether the information is relevant or not. Given the user’s known perception of higher-ranked rows as more relevant, upper rows are usually examined more extensively (Figure 1b).

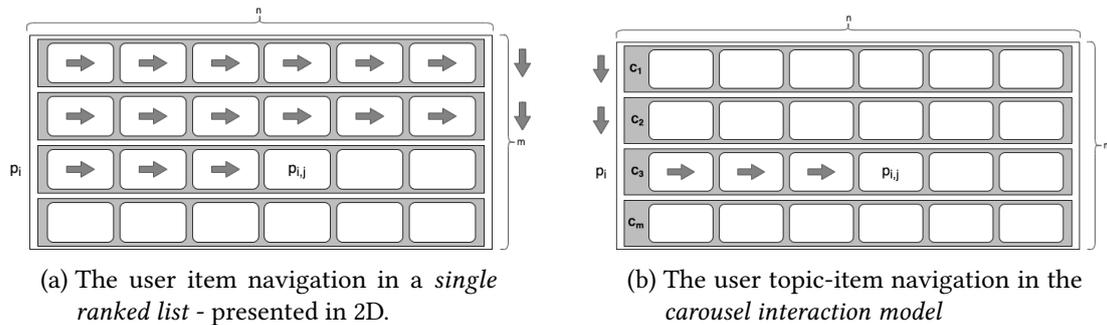


Figure 1: The schematic view of user navigation in *single ranked list* (left) and *carousel interaction model* (right).

The combination of these two factors created a characteristic “golden triangle” pattern in eye-tracking studies [41]. The results of studies of user behavior in search context helped to create the first 2D click models of user behavior [42, 40]. While no rigorous studies specifically focused on user work with carousel-based interfaces have been performed yet, the analysis of user behavior data obtained in recent studies of carousel-based interfaces [38, 39] demonstrated the same general golden triangle pattern. Based on this empirical data, we developed a model of user behavior in carousel-based recommender interfaces. This model is presented in a formal way in [43]. Below we present this model less formally to the extent it is necessary to understand how our simulation-based study was performed.

3.2.1. Modeling Assumptions

To quantify the benefit of carousels, we formalize the problem of carousel recommendation using a mathematical model, which we call a *carousel interaction model*. We have a matrix of $m \times n$ recommended items, where m is the number of rows (carousels) and n is the number of columns (items per carousel). Each carousel is associated with some topic, such as a movie genre. To simplify exposition, we assume that each item belongs to a single topic. We refer to the item at row $i \in [m]$ and column $j \in [n]$ as (i, j) .

The user preferences are defined by two sets of probabilities. The first is *topic preferences*. Specifically, $p_i \geq 0$ is the probability that the user is interested in topic i , for any $i \in [m]$. The second set is *topic-conditioned item preferences*. Specifically, $p_{j|i} \geq 0$ is the conditional probability that the user is interested in item j given that they desire topic i , for any $i \in [m]$ and $j \in [n]$. We assume that $\sum_{i=1}^m p_i = 1$, and that $\sum_{j=1}^n p_{j|i} = 1$ for any topic $i \in [m]$.

3.2.2. Simple Behavior Model

Based on the above assumptions, we developed three increasingly more complex behavior models. The simplest of these models assumes a patient and focused user who continues to examine topics and items until finds the desirable item, i.e., an item that matches user preferences and interests in the given moment. This user interacts in the carousel model as follows. First, we assume that the user starts with some understanding what kind of topic (the desired topic) and item in that topic (a desired item) she wants to consume before she starts using the recommender system interface to locate such an item. To simulate this process in a personalized way, the *desired topic* is sampled as $I \sim \text{Cat}((p_i)_{i=1}^m)$ and the *desired item* is sampled as $J \sim \text{Cat}((p_{j|I})_{j=1}^n)$, where $\text{Cat}(\theta)$ is a categorical distribution with outcome probabilities θ . In plain English, exactly one topic is chosen with probability p_i , and exactly one item is chosen with probability $p_{j|I}$ conditioned on that topic. An equivalent way of thinking of this process is that exactly one (i, j) is chosen with probability $p_{i,j} = p_{j|i}p_i$. The user seeks item (I, J) as follows. They start by examining the first carousel. If its topic does not match that of I , they proceed to the next carousel. The user examines all carousels, from top to bottom, until they stop at carousel I . After that, the user examines the items in carousel I , from left to right, until they find the desired item, in column J .

There are two different scenarios under which the user might leave the system. The session may end after the user successfully find the desired item or because none of the items are desirable for the user and there is no more item and topic to examine. We are aware that this browsing behavior is unlikely to occur in a realistic situation due to the position bias effect [30]. However, we start with this simple model to highlight the difference between this and other more realistic behavioral models.

3.2.3. Impatient User Model

The simple model assume that the user is patiently scanning presented items until the desirable item is found - even if it requires to go through thousands of items. In the majority of cases, this assumption is not realistic. To better model a browsing behavior of an actual user, we assume that the user has limited patience for finding the desirable item. We implemented this behavior

as follows. The user starts by examining the first topic or item at position $(1, 1)$. The user exits with a probability of $p_q = 0.02$ after examining either a carousel or item. Generally, users are likely to abandon the session after 50 interactions on average, when no items or topics are desirable. This is the same as the ideal setting except for exiting with probability $p_q = 0.02$ upon each examination, of either a carousel or an item.

3.2.4. Distracted User Model

We initially assumed that the user always knew which carousel (with a genre as a topic) includes the desirable movie. However, in reality, the user might get distracted and as a result, begin browsing the wrong carousel or pass the correct carousel and miss out on finding the desired item. We consider this assumption to be an extension of the previous assumption described in Section 3.2.3.

In both ideal and distracted user settings, when the user examines an undesirable carousel, they will move to the next carousel with a probability 1. We define $p_d = 0.05$ as the distraction probability. Here user moves to the next carousel with probability $1 - p_d$ and starts examining items in the undesirable carousel with probability p_d . Similarly, when the user examines a desirable carousel, they move to the next carousel with probability p_d and start examining items in the desirable carousel with probability $1 - p_d$. Considering a user as *distracted* only applies in *carousel interaction model*. Including this assumption in the *carousel interaction model* allows us to capture the complexity that comes with providing additional information to the user in the form of carousel topics.

Because of lacking a large enough data set that can accurately estimate the parameters of our proposed settings, we set the values of p_q and p_d intuitively based on how we presume the user would behave under those settings.

4. Carousels Versus Ranked List: A Simulation-Based Study

We conduct a series of offline simulation-based experiments to evaluate how our proposed *carousel interaction model* performs against a standard baseline (*single ranked list*). In this section, we discuss the details of our simulation-based study.

4.1. Dataset and Setting

For our experiments, we choose the domain of movie recommendation. The choice of the domain was motivated by two reasons. First, movie recommendation is a good example of a modern context where users can have multiple interests and favor different interests at different times. Second, it is the context where carousels are currently very popular, which makes it easier to simulate realistic carousel-based recommendations. We use the MovieLens 100K Dataset [44] which consists of 100,836 ratings applied to 9,724 movies in 19 genres by 610 users. In our experiments, we only utilize the information about the user ratings and movie genres. We apply a pre-processing step to remove movies with no genres. A total number of 34 movies was removed from the dataset through this process.

We assume that the user follows three distinct browsing behavior models when seeking a movie to watch depending on whether the results are presented as a single ranked list or a set of carousels. These three models are explained in Section 3.2.2, 3.2.3 and 3.2.4 respectively.

To generate the recommendations, we consider two sets of probabilities. The *topic preferences* and the *topic-conditioned item preferences*. The preferences are computed as follows. The dataset of ratings is a set of tuples $\mathcal{D} = \{(k_t, j_t, r_t)\}_{t=1}^n$, where k_t is the index of the user in data point t , j_t is the index of the rated movie in data point t , and r_t is the corresponding rating. The topic-conditioned item preference reflects how representative the movie is of a genre. We compute it as the sum of all ratings of the movie over the sum of all ratings in its genre. Formally, let \mathcal{G}_i be the set of all movies in genre i . Then for any movie $j \in \mathcal{G}_i$, the topic-conditioned item preference of movie j in genre i is

$$p_{j|i} = \frac{\sum_{t=1}^n \mathbb{1}\{j_t = j\} r_t}{\sum_{t=1}^n \mathbb{1}\{j_t \in \mathcal{G}_i\} r_t}.$$

We set $p_{j|i} = 0$ for any $j \notin \mathcal{G}_i$. For any user k , the topic preference reflects how much the user prefers a genre. We compute it as the sum of all ratings of the user in a given genre over all ratings by that user. Formally, the topic preference of user k for genre i is

$$p_i = \frac{\sum_{t=1}^n \mathbb{1}\{k_t = k, j_t \in \mathcal{G}_i\} r_t}{\sum_{t=1}^n \mathbb{1}\{k_t = k\} r_t}.$$

4.2. Recommendation Approach

Having the *user profile* assigned to each user, we generate two sets of recommendations as follows: For the first set of recommendations for carousels, we use the *topic preferences* to sort them and then populate each one with movies using the topic-conditioned item preferences. This approach generates a set of carousels each representing a genre (19 carousels for 19 genres in the dataset). Each carousel contains all the movies within the representative genre. With an average of more than 475 movies in each genre, we assume that is a realistic enough scenario for the user to be able to scroll down or right and examine all items and find the desirable movie. The movie is sorted by their scores, where the score of movie j is $\sum_{i=1}^m p_{i,j}$. Due to the sheer number of movies in the dataset, we assume that users will be able to scroll down the list to find what they are looking for. In this evaluation, the *user profile* and the recommendations were not affected by further user interactions and remained unchanged throughout all sessions.

4.3. Simulation Process

We define a session as a single instance of evaluation in which the user seeks a movie (I, J) from the set of recommended results, which can be displayed as a *single ranked list* or *carousel interaction model*.

The process of simulation is as follows: For each setting, we first generate two sets of recommendations (one using *single ranked list* and another using *carousel interaction model*) for every user in the dataset. Next, we ran 100 independent sessions for every user that includes selecting a genre, selecting a movie within that genre, and calculating the number of interactions

required to reach that movie in both models. We consider the average value of these 100 sessions as the outcome of the experiment for a given user in a given setting.

To simulate user navigation in each session, we assume that the desired genre and a movie in that genre are realized in the mind of the user. The *desired genre* is sampled as $I \sim \text{Cat}((p_i)_{i=1}^m)$ and the *desired movie* is sampled as $J \sim \text{Cat}((p_{j|I})_{j=1}^n)$. This process is described in detail in Section 3.2. In each session, the user is only interested in a single genre and a single movie within that genre.

4.4. Evaluation Metrics

There are many ways of measuring the complexity of interacting with the recommended items in *single ranked list* and *carousel interaction model*. We employ two metrics to evaluate our proposed approach.

First, we define *navigation effort* as the number of examinations by users until the desired item is found. These examinations include browsing genres as carousel topics and movies as items. A lower *navigation effort* means less effort to find a desirable item.

Second, we define the *exiting Probability* which determines on average what proportion of users left the session after a certain number of interactions. For example, in Figure 3a on average, just under 50% of users in *carousel interaction model* exited the session with fewer than 30 interactions. The total number of interactions includes all examinations done by the user to find the desirable movie. It is important to state that the *exiting Probability* only can be considered as a positive metric under the ideal setting where the user continues the examination until finding the desirable items.

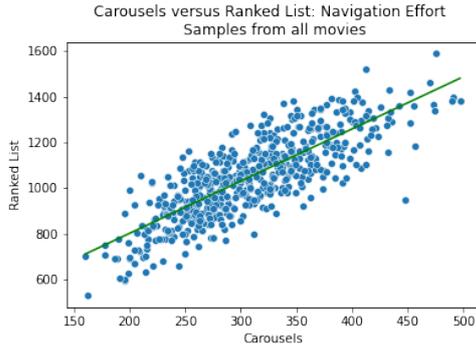
Unlike the ideal setting, in distracted and impatient settings, the *exiting Probability* could be an indication of either satisfactory, due to finding the desirable items or unsatisfactory, due to impatience or distraction and without necessarily finding the desirable items. In our experiments, we only compare the *exiting Probability* under the comparable settings.

5. Results

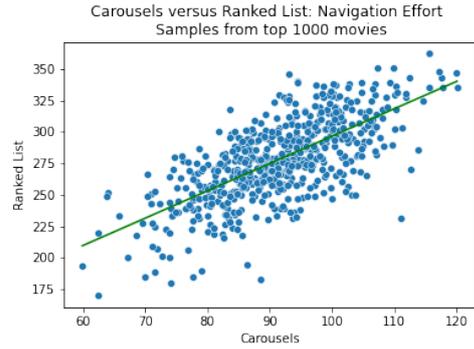
In this section, we present the results of our simulation-based evaluation. In Section 5.1 we compare the distribution of *navigation effort* for *carousel interaction model* and *single ranked list* models with samples from a different number of top movies in the dataset. In Section 5.2 we demonstrate how different browsing behavior affects the exiting pattern among users.

5.1. Navigation Effort

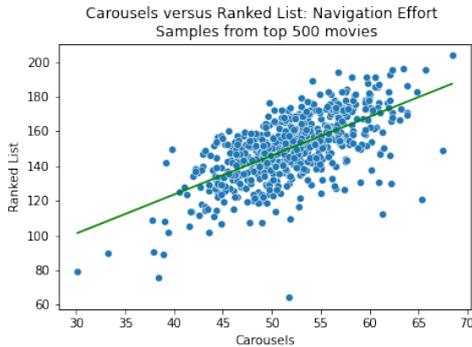
Figure 2a shows the distribution of *navigation effort* values for all the movies (9708) and users (610) in our dataset. The experiment was conducted under the ideal setting described in Section 3.2.2. Similarly, Figures 2b to 2d display the distribution of *navigation effort* values for the top 1000, 500 and 100 movies respectively. These results indicate that the *carousel interaction model* significantly reduced the number of required interactions to find a desirable movie. It is evident that even though the slope of the lines remains relatively steady, the higher number of



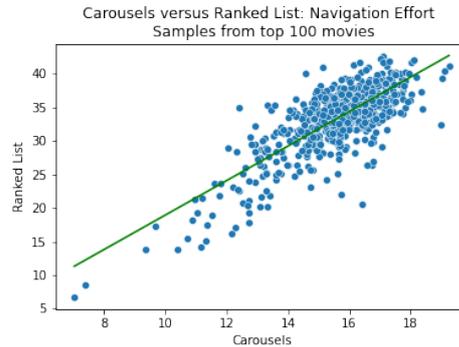
(a) Samples drawn from all movies.



(b) Samples drawn from top 1000 movies.



(c) Samples drawn from top 500 movies.



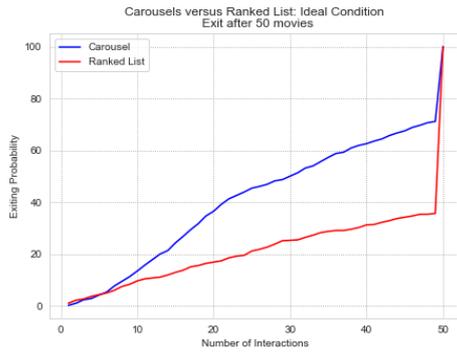
(d) Samples drawn from top 100 movies.

Figure 2: The correlation between navigation effort for a selected movie in *carousels* and *ranked list* models.

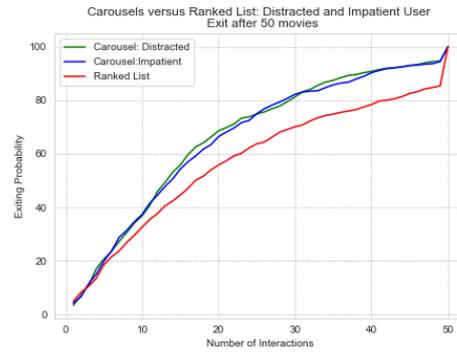
movies resulted in a slightly more prominent improvement from *single ranked list* to a *carousel interaction model*.

5.2. Exiting Probability

To compare the behavior of our model under more realistic settings, we visualize the average *exiting probability* of users after a certain number of interactions with the recommendations in Figure 3a and Figure 3b. In Figure 3a we observe a significant difference between the *carousel interaction model* and the *single ranked list* under the ideal settings. In the ideal setting, the user continues the examination until reaches the desirable item. We limit the number of interactions to 50 meaning the user would exit unsatisfied if they could not find the desired item in the first 50 interactions. The higher *exiting probability* in *carousel interaction model* (blue line) shows that more users exit the system satisfied by finding their desired item. A larger spike in *exiting probability* on *single ranked list* at the end indicates a larger number of users that left without finding their desired item. It is worth noting that based on the result of this experiment, a significantly larger portion of users (just under 75%) exit the system after finding their desirable. This number drops to close to 30% when recommendations are presented in the form of a



(a) Cumulative Exiting Probability under ideal condition.



(b) Cumulative Exiting Probability for *distracted* and *impatient* user.

Figure 3: Comparing the cumulative *exiting probability* in *single ranked list* and *carousel interaction model* and in different experimental settings reveals the advantages of using a carousel based representation compared to a ranked list based representation. In all experimental settings users leave after 50 interactions regardless of success in finding the desirable item.

ranked list. The exiting behaviour of the simulated *impatient* and *distracted* users is displayed in Figure 3b.

Although the gap between the probability of exiting the session in *carousel interaction model* and *single ranked list* models is less significant, the former still performs better. Comparing the *Impatient* and *distracted* exiting behavior indicates a non-significant difference between the two settings but shows a slight decrease in performance in *carousel interaction model*. Unlike Figure 3a where the *exiting probability* promote a positive event (satisfaction of finding the desirable item), in Figure 3b there can be also adverse reasons for exiting a session, such as "impatience" and "distraction". Therefore, the improvement of this metric compared to the ideal setting is not necessarily a positive sign. Despite this, since we compare *carousel interaction model* and *single ranked list* in Figure 3b under the same setting where the probability of "impatience" is the same, an improvement in the metric likely signal a positive event.

6. Conclusions

This paper makes a case for using simulation-based approaches for offline data driven evaluation of interactive recommender systems. We believe that the ability to use offline evaluation will benefit the research on interactive recommender systems in the same way as data-driven offline studies boosted the work on recommendation algorithms. To demonstrate the value and the opportunities for simulation-based evaluation we presented an example of a simulation-based study, which compared user behavior in a ranked list and multi-list interfaces. This study was enabled by empirically based models of user interaction with these two kinds of interfaces. The results of our comparative evaluation demonstrate the navigational superiority of the carousel-based interface and uncovers the reasons for its increasing popularity. While this result is important by itself, in the context of our paper it serves as an illustration of interesting

findings that could be made by using simulation-based evaluation.

The case presented in the paper could be considered as relatively simple - the benefits of carousel-based interfaces over simple ranked lists might be intuitively evident. Yet even in this simple case only with a thorough study these benefits can be measured and quantified. A simulation-based approach enables us to quantify differences through a fair comparative study and without engaging expensive human subjects. In a similar way, this approach could be applied to more complex scenarios where simple intuition will not suffice. We hope that this paper will help to promote simulation-based evaluation of all kinds of interactive recommenders. In our future work, we plan to continue exploration of carousel-based interfaces from the prospect of human-AI collaboration, explore more powerful approaches to ranking items within each topic-based carousel, and compare these approaches through simulation-based and empirical studies.

Acknowledgement

We would like to thank Branislav Kveton, for his great help with formalizing the ranked list and carousel interactions using mathematical models. His willingness to give his time so generously has been very much appreciated.

References

- [1] D. Chin, Empirical evaluations of user models and user-adapted systems, *User Modeling and User-Adapted Interaction* 11 (2001) 181–194.
- [2] Z. A. Pardos, N. Heffernan, KT-IDEM: introducing item difficulty to the knowledge tracing model, in: J. Konstan, R. Conejo, J. Marzo, N. Oliver (Eds.), *19th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2011*, Springer, 2011, pp. 243–254.
- [3] M. Desmarais, R. S. Baker, A review of recent advances in learner and skill modeling in intelligent learning environments, *User Modeling and User-Adapted Interaction* 22 (2012).
- [4] D. Kluver, M. Ekstrand, J. Konstan, *Rating-Based Collaborative Filtering: Algorithms and Evaluation*, LNCS, Springer, Heidelberg, 2017, pp. 344–390.
- [5] A. Gunawardana, G. Shani, Evaluating recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), *Recommender Systems Handbook*, Springer, Boston, MA, 2015.
- [6] M. Rossetti, F. Stella, M. Zanker, Contrasting offline and online results when evaluating recommendation algorithms, in: *Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, 2016, pp. 31–34.
- [7] L. Peska, P. Vojtas, Off-line vs. on-line evaluation of recommender systems in small e-commerce, in: *31st ACM Conference on Hypertext and Social Media, 2020*, pp. 291–300.
- [8] F. Ricci, L. Rokach, B. Shapira, Recommender systems: Introduction and challenges, in: *Recommender Systems Handbook*, Springer, Boston, MA, 2015, pp. 1–34.
- [9] C. He, D. Parra, K. Verbert, Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities, *Expert Systems with Applications* 56 (2016) 59–27.

- [10] L. Granka, T. Joachims, G. Gay, Eye-tracking analysis of user behavior in www search, in: 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 478–479.
- [11] Z. Guan, E. Cutrell, An eye tracking study of the effect of target rank on web search, in: CHI '07: ACM SIGCHI conference on human factors in computing systems, ACM Press, 2007, pp. 417–420.
- [12] K. Jarvelin, J. Kekalainen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446.
- [13] N. Craswell, Mean reciprocal rank, in: L. Liu, M. T. Özsu (Eds.), Encyclopedia of Database Systems, Springer US, Boston, MA, 2009, pp. 1703–1703.
- [14] P. Pirolli, W.-T. Fu, SNIF-ACT: A model of information foraging on the world wide web, in: P. Brusilovsky, A. Corbett, F. d. Rosis (Eds.), 9th International User Modeling Conference, volume 2702 of *Lecture Notes in Artificial Intelligence*, Springer Verlag, 2003, pp. 45–54.
- [15] J. Champaign, R. Cohen, Ecological content sequencing: From simulated students to an effective user study, International Journal of Learning Technology 8 (2013) 337–361.
- [16] D. Dzyabura, A. Tuzhilin, Not by search alone, in: 7th ACM conference on Recommender systems, 2013, p. 371–374.
- [17] N. Hazrati, M. Elahi, F. Ricci, Simulating the impact of recommender systems on the evolution of collective users' choices, in: 31st ACM Conference on Hypertext and Social Media, ACM, 2020, pp. 207–212.
- [18] J. Mostafa, S. Mukhopadhyay, M. Palakal, Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering, Information Retrieval 6 (2003) 199–223.
- [19] S. K. Card, P. Pirolli, M. Van De Wege, J. B. Morrison, R. W. Reeder, P. K. Schraedley, J. Boshart, Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability, in: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'2001), ACM Press, 2001, p. 498–505.
- [20] 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, in: FATES 2020 : 2nd Workshop on Fairness, Accountability, Transparency, Ethics and Society on the Web at The Web Conference 2020, 2020.
- [21] H. Drachsler, H. Hummel, R. Koper, Using simulations to evaluate the effects of recommender systems for learners in informal learning networks, in: R. Vuorikari, B. Kieslinger, R. Klamka, E. Duval (Eds.), 2nd SIRTEL'08 Workshop on Social Information Retrieval for Technology Enhanced Learning, volume 382 of *CEUR Workshop Proceedings*, CEUR, 2008.
- [22] N. Hazrati, F. Ricci, Simulating users' interactions with recommender systems, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, ACM, 2022, pp. 95–98.
- [23] D. Bountouridis, J. Harambam, M. Makhortykh, M. Marrero, N. Tintarev, C. Hauff, SIREN, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 150–159.
- [24] T. Joachims, Optimizing search engines using clickthrough data, in: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 133–142.
- [25] B. Pan, H. Hembrooke, G. Gay, L. Granka, M. Feusner, J. Newman, The determinants of

- web page viewing behavior: an eye-tracking study, in: ETRA'2004: Proceedings of the symposium on Eye tracking research and applications, ACM, 2004, pp. 147–154.
- [26] Z. Guan, E. Cutrell, What are you looking for?: An eye-tracking study of information usage in web search, in: CHI '07: ACM SIGCHI conference on human factors in computing systems, 2007, pp. 407–416.
- [27] M. Keane, M. O'Brien, B. Smyth, Are people biased in their use of search engines?, *Commun. ACM* 51 (2008) 49–52.
- [28] E. Agichtein, E. Brill, S. Dumais, R. Ragno, Learning user interaction models for predicting web search result preferences, in: 29th Annual International ACM SIGIR Conference, 2006, pp. 3–10.
- [29] M. Richardson, E. Dominowska, R. Ragno, Predicting clicks: Estimating the click-through rate for new ads, in: 16th International Conference on World Wide Web, 2007, pp. 521–530.
- [30] N. Craswell, O. Zoeter, M. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: Proceedings of the 1st ACM International Conference on Web Search and Data Mining, 2008, pp. 87–94.
- [31] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. M. Wang, C. Faloutsos, Click chain model in web search, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 11–20.
- [32] O. Chapelle, Y. Zhang, A dynamic Bayesian network click model for web search ranking, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 1–10.
- [33] S. Li, Y. Abbasi-Yadkori, B. Kveton, S. Muthukrishnan, V. Vinay, Z. Wen, Offline evaluation of ranking policies with click models, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, pp. 1685–1694.
- [34] A. Chuklin, I. Markov, M. de Rijke, *Click Models for Web Search*, Morgan & Claypool Publishers, 2015.
- [35] B. Kveton, C. Szepesvari, Z. Wen, A. Ashkan, Cascading bandits: Learning to rank in the cascade model, in: International Conference on Machine Learning, 2015.
- [36] R. Combes, S. Magureanu, A. Proutiere, C. Laroche, Learning to rank: Regret lower bounds and efficient algorithms, in: Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2015.
- [37] N. Felicioni, M. F. Dacrema, P. Cremonesi, A methodology for the offline evaluation of recommender systems in a user interface with multiple carousels, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, ACM, 2021, pp. 10–15.
- [38] D. Jannach, M. Jesse, M. Jugovac, C. Trattner, Exploring multi-list user interfaces for similar-item recommendations, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, ACM, 2021, pp. 224–228.
- [39] A. Starke, E. Asotic, C. Trattner, “Serving Each User”: Supporting different eating goals through a multi-list recommender interface, in: Fifteenth ACM Conference on Recommender Systems, ACM, 2021, pp. 124–132.
- [40] L. Wu, M. Grbovic, J. Li, Toward user engagement optimization in 2d presentation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, ACM, 2021, pp. 1047–1055.
- [41] F. Chierichetti, R. Kumar, P. Raghavan, Optimizing two-dimensional search results presen-

- tation, in: Proceedings of the 4th ACM International Conference on Web Search and Data Mining - WSDM '11, ACM Press, 2011, pp. 257–266.
- [42] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, Q. Mei, Beyond ranking: Optimizing whole-page presentation, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, ACM, 2016, pp. 103–112.
- [43] B. Rahdari, B. Kveton, P. Brusilovsky, The magic of carousels: Single vs. multi-list recommender systems, in: 33rd ACM Conference on Hypertext and Social Media, 2022, pp. 166–174.
- [44] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Transactions on Interactive Intelligent Systems 5 (2015) 1–19.