The Magic of Carousels: Single vs. Multi-List Recommender Systems

BEHNAM RAHDARI, University of Pittsburgh, USA
BRANISLAV KVETON, Amazon, USA
PETER BRUSILOVSKY, University of Pittsburgh, USA

Carousel-based interfaces with multiple topic-focused item lists have emerged as a de-facto standard for presenting recommendation results to end-users in real-life recommender systems. In this paper, we attempt to formalize and explain the "magic" power of carousel-based interfaces from a traditional hypertext prospect of navigability. By applying both, formal analysis and a data-driven evaluation, we demonstrate and measure the benefits offered by the carousel-based organization of recommendations. We hope that this work will benefit the researchers in both hypertext and recommender systems communities, where the research on carousel-based interfaces is gaining popularity.

CCS Concepts: • **Information systems → Recommender systems**; • **Human-centered computing → Interactive systems and tools**.

Additional Key Words and Phrases: Recommender Systems; Carousel-based interface; Human-AI collaboration, Navigability, Click Models

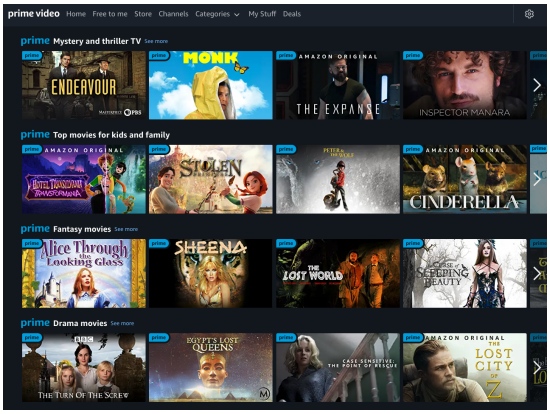
ACM Reference Format:
Behnam Rahdari, Branislav Kveton, and Peter Brusilovsky. 2022. The Magic of Carousels: Single vs. Multi-List Recommender Systems. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22), June 28-July 1, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3511095.3531278

1 INTRODUCTION

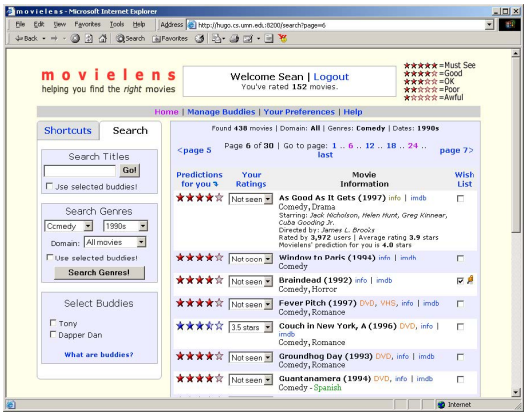
The field of recommender systems is now over 25 years old and accumulated a large body of research. The majority of this research has been focused on developing and evaluating more and more powerful ranking algorithms. Following the example of information retrieval systems, recommender systems strive to assess the relevance of candidate items and generate a "perfect" ranked list where the most relevant items are pushed to the top. Given this key goal, the power of a recommender system is evaluated by measuring its ability to estimate item relevance (i.e., predict item rating) and position it correctly in the ranked list. With so much attention to the ranked list as the outcome of the recommendation process, it might look like a surprise that in modern real-life recommender systems used by millions, it was not a traditional ranked list, but a two-dimensional structure in the form of a set of "carousels" (Figure 1a) that emerged as a de-facto standard to present recommendations to end-users.

While the interface with multiple carousels (sometimes referred to as a *multilist*) looks relatively complex – it presents several ranked lists, each marked with a category, in place of a single ranked list – it was embraced by the end-users and industry. This magical appeal of the carousels could be explained by at least two prospects. From the prospect of hypertext organization and navigability, a carousel-based interface offers a more powerful and better-structured

Manuscript submitted to ACM



(a) A carousel-based recommendation (Amazon Prime video, 2022).



(b) A ranked list recommendations (MovieLens v3, circa 2003).

Fig. 1. The majority of modern recommender systems currently use a set of carousels to present the results (left). Before the popularity of the carousels, recommendation results were presented similar to search results, in the form of a single ranked list (right, used from [22] with author's permission).

interface for user navigation than the traditional single ranked list (Figure 1b). In this aspect, it is similar to other structured and semantically-infused navigation approaches, such as faceted browsing. Faceted browsing allows the user to filter items by their attribute values across multiple attributes, and it is well known for navigational efficiency.

From the prospect of recommender systems, the carousel-based interface provides an excellent example of human-AI collaboration in recommendation context. While a single ranked list attempts to be "perfect", in reality the intent of the user is highly uncertain. Most importantly, in many real-life applications users might have multiple interests and recommender systems rarely know which specific interest (for example, a movie genre) the users want to pursue at the given moment. A carousel-based interface leaves the task of choosing the most timely topic of interest (i.e., British documentaries) to the users. As a result, a user could land directly on a ranked sub-list of most relevant items, while also indirectly informing the recommender system about the kind of items they prefer right now. Several types of "interactive" recommender interfaces where human and AI-based recommender system can collaborate in guiding users to the right items have been explored and their effectiveness have been convincingly demonstrated [6, 31, 35, 41].

The surprising popularity and power of carousel-based interfaces has not been ignored by the researchers on recommender systems. A growing number of papers focused on application and user evaluation of carousel-based interfaces have been published over the last few years [5, 19, 27, 36, 42]. Surprisingly, the power of these interfaces has not yet been explored from the prospects of information navigation. This paper attempts to bridge this gap by examining and explaining the "magic" power of carousel-based interfaces from both analytical and empirical prospects.

2 BACKGROUND

2.1 Human-AI collaboration in Recommender Systems

The problems of a single ranked list tuned to the "best overall" interests have been long recognized in research on recommender systems. Historically, most attempts to address this problem focused on producing a "better" one-shot list or recommendations rather than treating the recommendation process as a human-AI collaboration. One of the

oldest streams of related work is the research on increasing the *diversity* of recommendations [44]. This research was motivated by similar observations - stuffing the top of the recommended list with very similar items is not productive even though all these items have top relevance score. This research stream introduces diversity as one of the measures of recommendation quality and explored a range of approaches to diversify the ranked list [26, 39]. A more recent approach known as *context-aware recommendation* acknowledged that user's current preferences are defined not just by the overall interest modeled by traditional recommender systems, but also on immediate needs and interests defined by the context. However, instead of collaborating with the user in understanding this context, context-aware recommenders relied on AI alone to model both the overall interests and the context to produce a one-shot "context-adapted" ranked list [1, 2, 38].

In contrast, closely related streams of research on "conversational", "dialogue", "critique-based", and "interactive" recommendation specifically focused on human-AI collaboration. The oldest stream of this work suggested structuring the recommendation process as a human-machine natural language dialogue where a human provides feedback (known as *critique*) to each (and originally single) suggestion of a recommender system [9]. This feedback allows the system to better understand users' current preferences and gradually improve recommendations. More recent research generalized this critique-based approach and attempted to make it more efficient using graphical user interfaces (GUI) in place of the natural language dialogue [11]. The use of GUI made the interaction process more efficient by recommending a list of items at each step and enabling the users to contribute more information through a more complex *compound critiques* [35]. Modern interactive recommender systems [24] made the enhanced GUI a centerpiece of human-AI interaction. Instead of a turn-taking dialog with a recommender agent, these systems offered users an opportunity to tune the results of recommendation continuously using sliders and other forms of direct manipulation [4, 6, 31, 32, 41].

Carousel-based recommender interfaces offer an interesting compromise between these two streams of research. Carousel-based interfaces inherit serious attitude to ranking from the AI-focused stream of research on recommender systems. At the same time it follow the stream of work on human-AI interaction where each partner does what they could do best. It is left for the human user to choose the current topic(s) of interest (represented by one or more offered carousels), an easy task for the user but a really hard one for AI that even good context-aware recommenders can't do reliably. It is left for the AI component to rank the topics of interests and the items within each selected topic, the task where AI ability to carefully analyse user interaction with the system supersedes user abilities.

2.2 Navigability in Hypertext Systems

With the growth of size and complexity of hypertext networks, the research on *navigability* emerged as an important topic in the hypertext and Web research. The navigability research usually focuses on the ability (or probability) to reach specific designation(s) when navigating various kinds of hyperspaces with or without help of navigation artifacts. The first generation of research on navigability focused mostly on understanding the navigation properties of the Web as a novel information artifact by itself [7, 28]. The results of this research were critical to better understand the nature of the Web and to develop more efficient search and navigation approaches. With the gradual development of various navigation artifacts (structures) such as concept indexes for concept-based navigation [8, 12], semantic categories for faceted browsing [43], and tags for tag-based navigation [18] the navigability research refocused on studying the navigation properties or these artifacts and comparing different approaches to create these artifacts from navigability prospect. For example, Venetis et al. [40] compared navigability of different approaches to generate navigational tag clouds, Trattner et al. [37] compared navigability of regular and faceted tag clouds, and Helic et al. [25] examined the

effect of automatic linking on navigability. This second generation of navigability research plays an important role in developing more efficient types of navigation artifacts and finding best approaches to generate them.

To explore navigability, researchers use both theoretical and empirical approaches. For example, Chi and Mytkowicz [13] examined the navigability of tag clouds from information-theoretical prospects while Venetis et al. [40] performed this analysis empirically. In turn, empirical evaluation could be performed using data-driven simulation [34, 40] or through a user study [37, 43]. While a user study is considered the most realistic approach to evaluating navigability, the complexity of organizing large-scale user studies leads to the domination of data-driven empirical approaches [7, 18, 34, 40]. The key idea of data-driven evaluation is to simulate user navigation in the information space with or without navigation artifacts using realistic models of user behavior. Navigability of information spaces could be considered as the simplest case of data-driven evaluation since it typically uses simple behavior models such as random walks [7] and requires no artifact generation. Personalization and navigation artifacts introduce additional complexity to data-driven evaluation. Personalization requires engaging realistic user data, for example, user bookmarks were necessary to evaluate personalized PageRank [23]. Navigability evaluation of navigation artifacts requires building more complex behavior models to simulate user interaction with these artifacts and using real data to build realistic artifacts, for example, realistic tag clouds [34, 40].

In this work, we explore navigability of a carousel-based recommendation interface, a new example of navigation artifact in the information space, by comparing it with traditional ranked list. To perform a comprehensive evaluation, we use both an analytical algorithmic-theoretic approach and an empirical data-driven approach. Since we study navigability of personalized navigation artifacts, we have to build a realistic model of user carousel interaction and bring in real user data to generate realistic interfaces and simulate user behavior in a personalized way.

2.3 Click Models in Recommender Systems

Interaction of users with ranked lists of recommended items has been long studied under the name of click models [14]. Many click models exist [3, 10, 16, 20, 21, 33]. Essentially all of them try to explain the user behavior by a generative model, which can be learned from data. As an example, the cascade model [16, 33] 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. This information can be used to debias logged data [30], or to learn better ranking policies either offline [14] or online [15, 29]. In this work, we study a generative model of user behavior that explains why interactions with carousels can be more efficient than with a single ranked list.

3 APPROACH

To uncover the "magic" of the carousels, we compare user interaction 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.

To compare carousels with ranked lists, we use two complementary approaches. First, we propose a mathematical model of user interaction with carousels, and use it to compare user interaction with carousels and ranked lists *analytically*. This comparison is presented in Section 4. Next we conduct a series of experiments using real user data to compare user interaction with carousels and ranked lists *empirically* from hypertext *navigability* prospect. The design of the experiments and their results are presented in Section 5. Finally, we conclude and discuss potential future work in Section 6.

4 CAROUSELS VERSUS RANKED LIST: COMPLEXITY ANALYSIS

This section is structured as follows. In Section 4.1, we introduce our carousel interaction model. In Section 4.2, we describe how to measure the complexity of user interaction with carousels. In Section 4.3, we compare the complexity of interaction in a single ranked list with that in the carousels.

4.1 Carousel Interaction Model

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 are *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 are *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 $i \in [n]$.

The user interacts in the carousel model as follows. First, the desired topic and item in that topic are realized in the mind of the user, and then the user seeks them. In particular, the *desired topic* is sampled as $I \sim \mathrm{Cat}((p_i)_{i=1}^{m})$ and the *desired item* is sampled as $J \sim \mathrm{Cat}((p_{j|I})_{j=1}^{n})$, where $\mathrm{Cat}(\theta)$ 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. A flowchart of how the user interacts with the carousels is presented in Figure 2. The two *End of Session* nodes on the bottom left and right represent different scenarios under which the user might leave the system. The session may end after the user successfully find the desirable item (left) or because none of the items are desirable for the user and there is no more item and topic to examine (right).

4.2 Complexity of Interaction in a Carousel Model

The complexity of interacting with the carousels can be measured in many ways. We define it as the number of examined carousels and items until the desired item is found. Specifically, if the desired item is (I, J), the number of interactions is $I + J$. Therefore, the expected number of interactions of the user with the carousels is $\sum_{i=1}^{m} \sum_{j=1}^{n} p_{i,j}(i + j)$. To illustrate this formula, suppose that each item is desired with probability $p_{i,j} = 1/(mn)$, which can be attained by setting $p_i = 1/m$

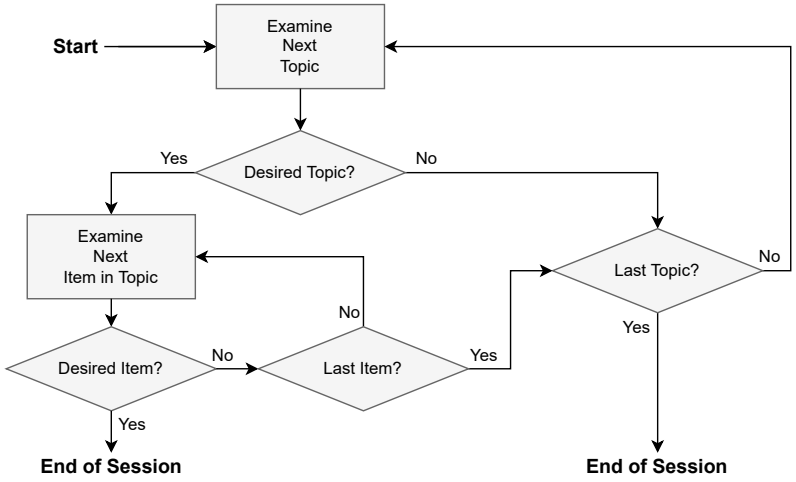


Fig. 2. Interaction between the user, topics, and items in the *carousel interaction model*.

and $p_{j|i} = 1/n$. Then the expected number of interactions is

$$\mathbb{E}\left[I + J\right] = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (i + j) = \frac{1}{mn} \left(n \sum_{i=1}^{m} i + m \sum_{j=1}^{n} j \right)$$

$$= \frac{1}{mn} \left(n \frac{m(m+1)}{2} + m \frac{n(n+1)}{2} \right) = \frac{m+n}{2} + 1 = O(m+n) \,.$$

This result is notable for the following reason. Although we recommend mn items, and they are equally popular, the expected number of interactions to find the desired item is only $O(m+n)$. This is because the items are organized and examined in a structured manner.

We would like to comment on two aspects of the model. First, the expected number of interactions is minimized when the carousels are sorted in descending order of p_i, and the items in each carousel are sorted in descending order of $p_{j|i}$. Second, p_i and $p_{j|i}$ are typically unknown and have to be estimated from data. We do that in our experiments in Section 5.

4.3 Complexity of Interaction in a Ranked List

To show improvements over a single ranked list, it is useful to think of the matrix of $m \times n$ recommended items in Section 4.1 as a single ranked list, which is examined row by row. 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 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. A flowchart of how the user interacts with a ranked list is presented in Figure 3.

Similarly to the carousel model, we measure the complexity of interaction by the number of examined items until the desired item is found. This means that the number of interactions to find item (i, j) is $n(i-1) + j$. This immediately leads to the following observation. For any item (i, j) where $i \geq 2$, the number of interactions in the carousel model,

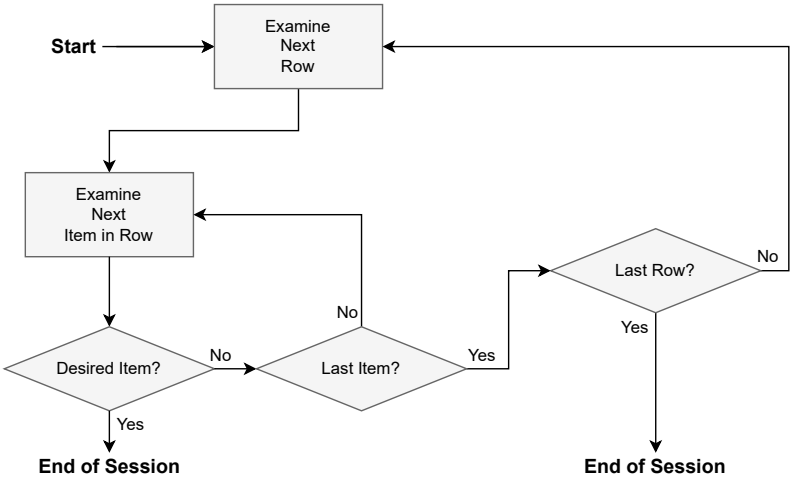


Fig. 3. Interaction between the user and items in a *single ranked list*.

$i + j$, is lower than in the single ranked list, $n(i-1)+j$, when the number of columns is $n \geq 2$. This is because

$$i + j \leq 2(i-1) + j \leq n(i-1) + j$$

when $i \geq 2$ and $n \geq 2$. The number of interactions for items in row 1 is higher in the carousel model because the user has to examine the carousel topic before it examines the items.

The above result suggests that the carousel model can lead to major improvements when most desired items are not concentrated in the first carousel. To show this, we consider the example from Section 4.1 where each item is desired with an equal probability $1/(mn)$. In this example, the expected number of interactions in the ranked list is

$$\mathbb{E}\left[n(I-1)+J\right] = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(n(i-1)+j)$$

$$= \frac{1}{mn}\sum_{\ell=1}^{mn}\ell = \frac{mn(mn+1)}{2mn} = \frac{mn+1}{2} = O(mn)\,.$$

In comparison, we get $O(m+n)$ interactions in the carousel model. Under the assumption that $m \approx n$, the savings in the number of interactions are the root of the number recommended items, $O(\sqrt{mn})$.

5 CAROUSELS VERSUS RANKED LIST: EXPERIMENTS

We conduct a series of data-driven experiments to evaluate how our proposed *carousel interaction model* performs against a standard baseline (*single ranked list*). These experiments complement our analytical evaluation by comparing the two recommendation interfaces in more realistic settings. 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.

This presentation below is structured as follows. In Section 5.1, we detail our experimental setup. In Section 5.2, we introduce different settings under which we evaluate our proposed model. Finally, we discuss our results in Section 5.3.

5.1 Setup

We use the MovieLens 100K Dataset [22] 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 adopts two distinct browsing behavior when seeking movie (I, J) providing that the results are presented as a single ranked list or a set of carousels. These two browsing behaviors are explained in Sections 4.1 and 4.3, 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

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

Having the *user profile* assigned to each user, we generate two sets of recommendations as follows: For the first set of recommendation 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 are 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 in 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.

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 4.1. In each session, the user is only interested in a single genre and a single movie within that genre.

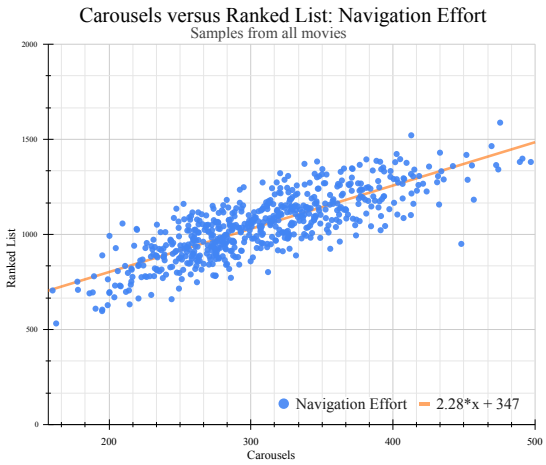
There are many ways of measuring the complexity of interacting with the recommended items in *single ranked list* and *carousel interaction model*. We employ two metrics to evaluate our proposed approach. First, we define *navigation effort* as the number of examinations by users until the desired item 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 5a 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.

5.2 Settings

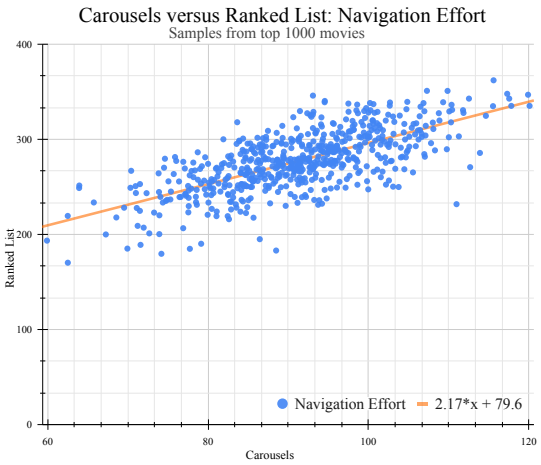
5.2.1 Ideal Setting. In the first setting, we assumed that the user continues to examine topics and items until finds the desirable item. The behavior of such a user is described in Section 4.1. We are aware that this browsing behavior is unlikely to occur in a realistic situation due to the position bias effect [17]. However, we include this setting in our evaluation to highlight the difference between this and other more realistic behavioral patterns.

5.2.2 Impatient User. 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.

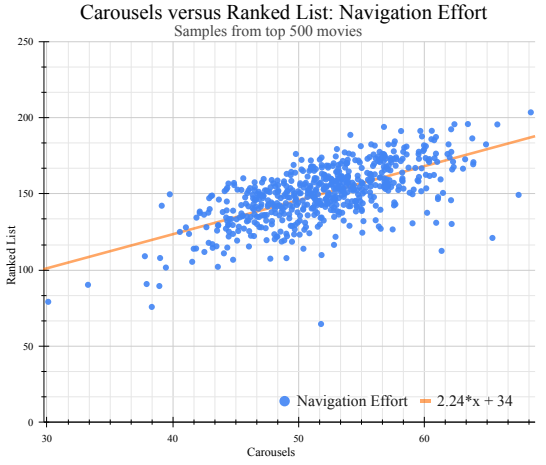
5.2.3 Distracted User. We initially assumed that the user always knew which carousel (with a genre as a topic) includes the desirable movie. However, in reality, the user might get distracted and as a result, begin browsing the wrong carousel or pass the correct carousel and miss out on finding the desired item. We consider this assumption to be an extension of the previous assumption described in Section 5.2.2. 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.



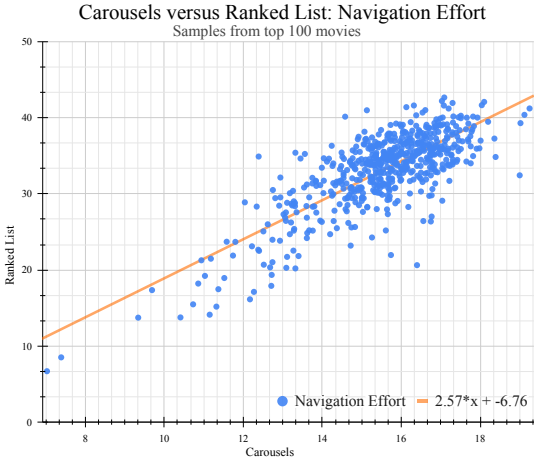
(a) Samples drawn from all movies.



(b) Samples drawn from top 1000 movies.



(c) Samples drawn from top 500 movies.



(d) Samples drawn from top 100 movies.

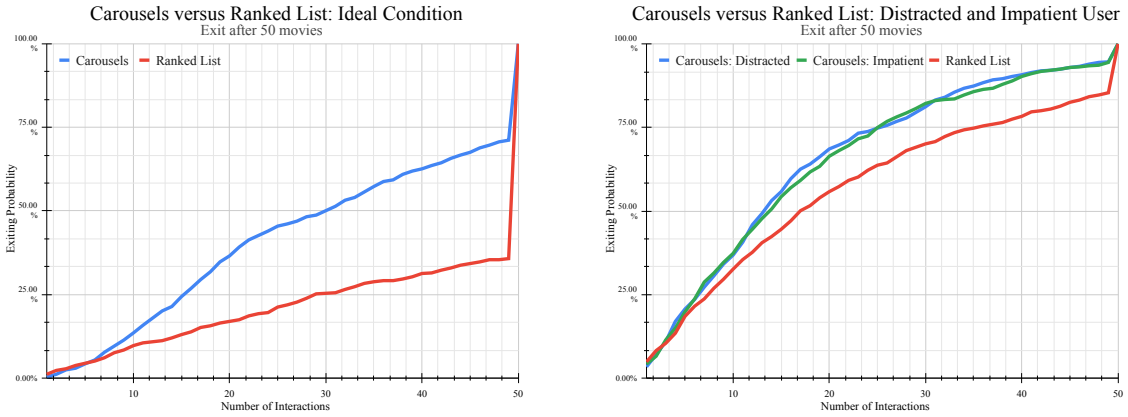
Fig. 4. The correlation between navigation effort for a selected movie in *carousels* and *ranked list* models.

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.

5.3 Results

In this section, we present the results of our experiments. In Section 5.3.1 we compare the distribution of *navigation effort* for *carousel interaction model* and *single ranked list* models with samples from different number of top movies in the dataset. In Section 5.3.2 we demonstrate how different browsing behavior affects the exiting pattern among users.

5.3.1 Navigation Effort. Figure 4a 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 5.2.1. Similarly,



(a) Cumulative Exiting Probability under *ideal* condition.

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

Fig. 5. 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.

Figures 4b to 4d 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 movies resulted in a slightly more prominent improvement from *single ranked list* to a *carousel interaction model*.

5.3.2 Exiting Probability. To compare the behavior of our model under more realistic settings, we visualize the average *exiting probability* of users after certain number of interactions with the recommendations in Figure 5a and Figure 5b. in Figure 5a we observe a significant difference between the *carousel interaction model* and the *single ranked list* under the ideal settings. In ideal setting, 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 desirable item in first 50 interactions. The higher *exiting probability* in *carousel interaction model* (blue line) shows that more users exit the system satisfied by finding their desirable 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 desirable item. It 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 ranked list. The exiting behaviour of the simulated *impatient* and *distracted* users is displayed in Figure 5b.

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 5a where the *exiting probability* promote a positive event (satisfaction of finding the desirable item), in Figure 5b 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 5b under the same setting where the probability of "impatience" is the same, an improvement in the metric likely signal a positive event.

6 CONCLUSIONS

In this paper we attempt to uncover the reasons for the rapidly increasing popularity of carousel-based recommendation interfaces (the magic of carousels) by comparing the navigability of carousel-based interfaces and a traditional *single ranked list*. We formalize the problem of carousel recommendation using a mathematical model called a *carousel interaction model*, and use this model to compare the two recommendation interfaces both analytically and empirically. To perform the analytical evaluation, we represented both interfaces in a comparable form as matrices and demonstrated that the use of topic-based carousels leads to a considerable reduction in the number of user interactions to find the desired item. Our analysis shows that the exact savings in interactions is equal to the square root of total number of recommendations which can be consider a significant improvement.

To support the analytical comparison, we compared the two interfaces in a sequence of increasingly more realistic settings using a data-driven empirical evaluation approach. Our experiments showed a considerable advantage of *carousel interaction model* over a *single ranked list* with respect to two evaluation criteria: *navigation effort* and *exiting probability*.

Taking together, the results of our comparative evaluation of *carousel interaction model* and *single ranked list* demonstrate the navigational superiority of the carousel-based interface and uncovers the reasons for its increased popularity. While this result is important by itself, we believe that our work also contributes to the research on navigability of carousel-based interfaces by offering a formalized *carousel interaction model* and demonstrating two principal approaches that could be used for evaluating more complex types of carousel-based interfaces.

In our future work, we plan study carousel-based interfaces from the prospect of human-AI collaboration, expanding current work in two directions. First, we would like to explore more powerful approaches to ranking items within each topic-based carousel. Second, we intend to augment data-driven empirical evaluation with a user-based evaluation, which will also help us to understand how users interact with carousel-based interfaces and build better interaction models.

REFERENCES

[1] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alexander Tuzhilin. 2011. Context-Aware Recommender Systems. *AI Magazine* 32, 3 (2011), 67–80.

[2] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Transactions on Information Systems.* 23, 1 (2005), 103–145.

[3] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. 2006. Learning User Interaction Models for Predicting Web Search Result Preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference.* Association for Computing Machinery, 3–10.

[4] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, Rene Witte, Greg Butler, and Adrian Tsang. 2013. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. In *international conference on Intelligent user interfaces, IUI '2013.* ACM Press, 49–56.

[5] Walid Bendada, Guillaume Salha, and Bontempelli. 2020. Carousel Personalization in Music Streaming Apps with Contextual Bandits. In *Fourteenth ACM Conference on Recommender Systems.* ACM, 420–425.

[6] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *6th ACM Conference on Recommender System.* 35–42.

[7] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual (Web) search engine. In *Seventh International World Wide Web Conference*, Helen Ashman and Paul Thistewaite (Eds.), Vol. 30. Elsevier Science B. V., 107–117.

[8] Peter Brusilovsky and Elmar Schwarz. 1997. Concept-based navigation in educational hypermedia and its implementation on WWW. In *ED-MEDIA/ED-TELECOM'97 - World Conference on Educational Multimedia/Hypermedia and World Conference on Educational Telecommunications*,

Tomasz Müldner and Thomas C. Reeves (Eds.). AACE, 112–117.

[9] Robin Burke, Kristian Hammond, and Benjamin C. Young. 1997. The FindMe Approach to Assisted Browsing. *IEEE Intelligent Systems* 12, 4 (1997), 32—40.

[10] Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the 18th International Conference on World Wide Web*. 1–10.

[11] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 125–150.

[12] Ed H. Chi, Lichan Hong, Julie Heiser, and Stuart K. Card. 2006. ScentIndex: Conceptually Reorganizing Subject Indexes for Reading. In *IEEE Symposium on Visual Analytics Science and Technology, VAST 2006*, Pak Chung Wong and Daniel Keim (Eds.). IEEE, 159–166.

[13] Ed H. Chi and Todd Mytkowicz. 2008. Understanding the Efficiency of Social Tagging Systems using Information Theory. In *The 19th ACM Conference on Hypertext and Hypermedia*. 81–88.

[14] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.

[15] Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. 2015. 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*.

[16] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*. 87–94.

[17] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) *(WSDM '08)*. Association for Computing Machinery, New York, NY, USA, 87–94. https://doi.org/10.1145/1341531.1341545

[18] Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. 2017. *Tag-based navigation and visualization*. LNCS, Vol. 10100. Springer, Heidelberg, in this volume.

[19] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. 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, 10–15.

[20] Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael Taylor, Yi Min Wang, and Christos Faloutsos. 2009. Click Chain Model in Web Search. In *Proceedings of the 18th International Conference on World Wide Web*. 11–20.

[21] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient Multiple-Click Models in Web Search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*. 124–131.

[22] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages.

[23] Taher H. Haveliwala, Sepandar D. Kamvar, and Glen Jeh. 2003. *An Analytical Comparison of Approaches to Personalizing PageRank*. Technical Report 2003-35. Stanford University. http://dbpubs.stanford.edu/pub/2003-35

[24] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56, C (2016), 59–27.

[25] Denis Helic, Sebastian Wilhelm, Ilire Hasani-Mavriqi, and Markus Strohmaier. 2011. The Effects of Navigation Tools on the Navigability of Web-based Information Systems. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM, 16:1–16:8. https://doi.org/10.1145/2024288.2024308

[26] Tamas Jambor and Jun Wang. 2010. Optimizing Multiple Objectives in Collaborative Filtering. In *2010 ACM conference on Recommender systems, RecSys '10*. ACM, 55–62.

[27] Dietmar Jannach, Mathias Jesse, Michael Jugovac, and Christoph Trattner. 2021. Exploring Multi-List User Interfaces for Similar-Item Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 224–228.

[28] Jon Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (1999), 604–632.

[29] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. 2015. Cascading Bandits: Learning to Rank in the Cascade Model. In *Proceedings of the 32nd International Conference on Machine Learning*.

[30] Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S. Muthukrishnan, Vishwa Vinay, and Zheng Wen. 2018. Offline Evaluation of Ranking Policies with Click Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1685–1694.

[31] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[32] Behnam Rahdari, Peter Brusilovsky, and Dmitriy Babichenko. 2020. Personalizing Information Exploration with an Open User Model. In *31st ACM Conference on Hypertext and Social Media*. ACM, 167–176.

[33] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web*. 521–530.

[34] Dimitrios Skoutas and Mohammad Alrifai. 2011. Tag clouds revisited. In *20th ACM International Conference on Information and Knowledge Management, CIKM '11*. ACM, 221–230.

[35] Barry Smyth, Lorraine McGinty, James Reilly, and Kevin McCarthy. 2004. Compound Critiques for Conversational Recommender Systems. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. IEEE, 145–151.

[36] Alain Starke, Edis Asotic, and Christoph Trattner. 2021. "Serving Each User": Supporting Different Eating Goals Through a Multi-List Recommender Interface. In *Fifteenth ACM Conference on Recommender Systems*. ACM, 124–132.

[37] Christoph Trattner, Yi-Ling Lin, Denis Parra, Zhen Yue, William Real, and Peter Brusilovsky. 2012. Evaluating Tag-Based Information Access in Image Collections. In *Proceedings of the 23rd ACM conference on Hypertext and hypermedia*. ACM, New York, NY, USA, 113–122.

[38] Mark van Setten, Stanislav Pokraev, and Johan Koolwaaij. 2004. Context-aware recommendations in the mobile tourist application COMPASS. In *Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004) (Lecture Notes in Computer Science, Vol. 3137)*, Paul De Bra and Wolfgang Nejdl (Eds.). Springer-Verlag, 235–244.

[39] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *the Fifth ACM Conference on Recommender Systems*. ACM Press, 109–116.

[40] Petros Venetis, Georgia Koutrika, and Hector Garcia-Molina. 2011. On the selection of tags for tag cloudsx. In *Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, Vol. 29. ACM, 835–844.

[41] Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Transactions on Interactive Intelligent Systems* 2, 3 (2012), Article 13.

[42] Liang Wu, Mihajlo Grbovic, and Jundong Li. 2021. Toward User Engagement Optimization in 2D Presentation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, 1047–1055.

[43] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti A. Hearst. 2003. Faceted metadata for image search and browsing. In *ACM Conference on Human Factors in Computing Systems, CHI 2003*. ACM Press, 401–408. http://citeseer.ist.psu.edu/yee03faceted.html

[44] Cai N. Ziegler, Sean M. McNee, Joseph Konstan, and Georg Lausen. 2005. Improving Recommendation Lists Through Topic Diversification. In *14th international conference on World Wide Web, WWW'2005*. ACM Press, 25–32.