CARE: An Infrastructure for Evaluation of Carousel-Based Recommender Interfaces

Behnam Rahdari
University of Pittsburgh
Pittsburgh, PA, USA
ber58@pitt.edu

Peter Brusilovsky
University of Pittsburgh
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT

Carousel-based interfaces are integral in enhancing user experience in online recommender systems like streaming services or e-commerce platforms, yet their usability evaluation often lacks standardization. Existing work on evaluating recommender systems, from toolkits to infrastructure, mainly assesses recommendation algorithms rather than user experience. This focus leads to a limited understanding of recommender systems' effectiveness, as it overlooks the role of user interface design, especially carousel-based interfaces, in user experience. In response, this paper introduces a web-based infrastructure for the usability assessment of carousel-based interfaces. Our infrastructure is adaptable for various domains and setups, and its modular design allows for potential expansion.

ACM Reference Format:
Behnam Rahdari and Peter Brusilovsky. 2024. CARE: An Infrastructure for Evaluation of Carousel-Based Recommender Interfaces. In *29th International Conference on Intelligent User Interfaces - Companion (IUI Companion '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3640544.3645223

1 INTRODUCTION

Carousel-based interfaces have grown popular in recommender systems, dynamically presenting recommendations in thematically organized lists and transforming user-content interaction on digital platforms. Their role in enhancing user experience, particularly in domains prioritizing engagement and content discoverability, is significant [11]. Studies indicate that carousels provide an engaging, exploratory user experience, improving content discovery and interaction [14, 17].

Traditionally, recommender systems evaluation focused on algorithmic efficiency and accuracy. However, this has evolved to include user interaction and satisfaction [11, 19], acknowledging that system success also depends on usability and user experience [8]. Despite this, interface evaluation, especially for carousel-based systems, has been underemphasized. This oversight leads to a limited view of system effectiveness, as user interaction is key to user experience [10].

IUI Companion '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0509-0/24/03
https://doi.org/10.1145/3640544.3645223

Most evaluation frameworks and infrastructures focus on algorithmic performance, overlooking the importance of an "interface-first" approach for carousel interfaces [7]. Addressing this, our paper introduces CARE , a novel infrastructure for user-centered evaluation of carousel-based interfaces. We emphasize interface evaluation in recommender systems, offering a balanced approach to algorithmic and interface design evaluation. Our goal is to enhance understanding of recommender systems, prioritizing user experience in evaluation efforts.

2 RELATED WORK

Carousel-based interfaces are increasingly popular in recommender systems, enhancing navigation [17], discovery [16], and user engagement by offering recommendations in thematically organized lists. These interfaces, prominent in entertainment and other domains like healthcare and education [13, 15], encourage users to explore beyond immediately visible items, leading to diverse discoveries [11].

Evaluating recommender systems traditionally focused on algorithmic metrics. However, the need for user-centered evaluation, considering metrics like diversity and novelty, has been increasingly recognized [8, 9]. Reliable evaluation now requires encompassing both algorithms and interfaces [1, 2]. Yet, evaluating carousel interfaces poses challenges, such as understanding user navigation and interaction patterns, assessing information overload, and the usability of control mechanisms. The lack of standardized tools hinders effective evaluation [5].

Recent studies focus on both algorithmic and visual aspects of recommender systems. EvalRS [3] and Cornac [18] focus on algorithmic evaluations, while Elliot [1] and EasyStudy [6] assess user interfaces. Despite the growth in user-centric approaches and new frameworks, the emphasis on algorithms persists, as seen in projects like POPROX [4]. Our paper introduces an infrastructure designed for interface-first evaluation, addressing this gap.

3 CARE : A CAROUSEL EVALUATION INFRASTRUCTURE

CARE facilitates the design, execution, and analysis of experiments with carousel-based recommender system interfaces. It offers customization for study creation, including task assignment and interaction logging, for detailed data collection and analysis.

CARE supports diverse user studies with detailed interaction logging in a controlled environment, aiding interface evaluation tasks like guiding users to find items within the carousel, assuming organic user engagement.

The experiment creation interface, structured into sections, guides researchers through defining and customizing their studies

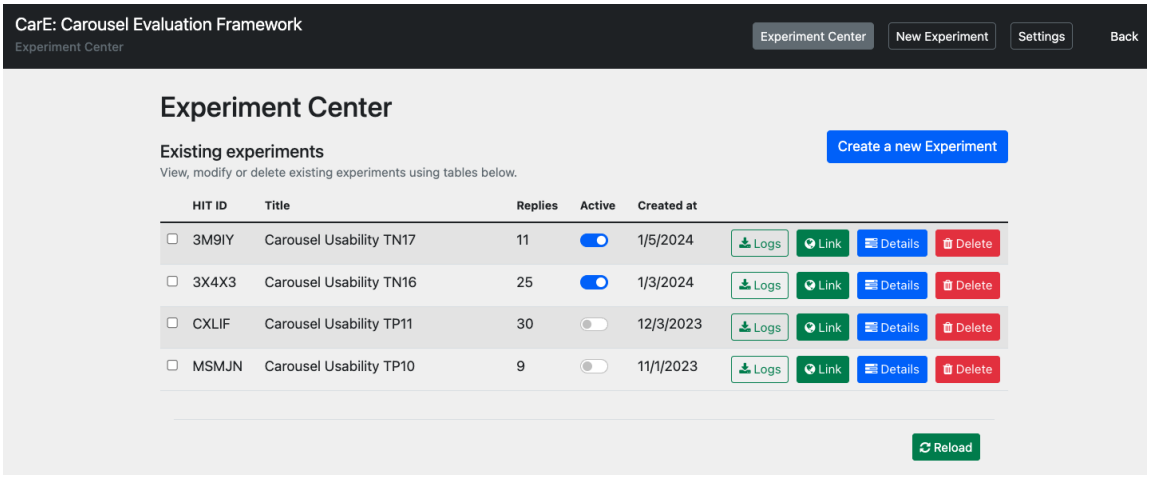

Figure 1: The view of the "Experiment Center" in CARE system.

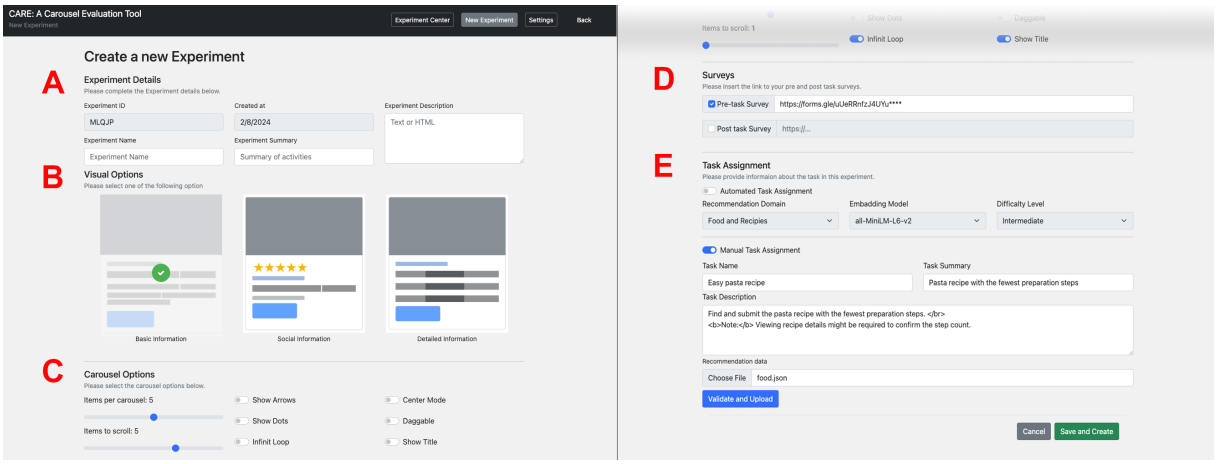

Figure 2: The interface for creating a new experiment in CARE system.

(Figure 2). The first section (Figure 2-A) allows entering general experiment metadata. Researchers can then select carousel themes and customize participant interaction settings (Figure 2-C), affecting usability and engagement.

Survey integration (Figure 2-D) links external surveys for qualitative data, while task assignment (Figure 2-E) offers automated and manual options. Automated assignment uses a pre-trained model for recommendations, while manual assignment uses a JSON file for detailed task control.

With the configuration complete, researchers launch experiments on platforms like Mechanical Turk, with streamlined MTurk integration. The Experiment Center monitors study progress and results (Figure 1).

The Experiment Center (Figure 1) displays ongoing experiments and participant data, allowing researchers to manage access and download interaction logs. Logging captures detailed user interactions like clicks and scrolls, adjustable for data granularity and storage management. This rich data aids in-depth behavioral analysis, which is discussed in the following section.

4 PROOF OF CONCEPT STUDY

We present an analysis of log data from a proof-of-concept experiment using CARE , primarily for data collection. The study involved participants finding the "lowest-rated sci-fi movie" in a custom set movie recommendations, with a carousel displaying four items and navigable via arrow keys.

Ten MTurk participants engaged in the study without demographic data collection but completed a pre-task survey. Each received a 50-cent compensation. Our logging infrastructure captured detailed mouse movement, scrolling, click data, and interacted UI element IDs and classes. Table 1 shows key user interaction metrics like experiment duration, mouse movement distances, and speeds.

Figure 3 displays two visualizations from our log data. The left panel shows a two-dimensional kernel density estimation (KDE) plot on a faded interface screenshot, indicating user click densities

Metric	Mean	Median	SD
Task Time (s)	26.53	23.21	12.31
Mouse Movement Distance (px)	2419.59	2002.20	1702.58
Mouse Speed (px/s)	184.80	176.92	93.57
Mouse Movement Events	480.3	400.8	215.1
Scroll Movement Events	85.40	62.00	63.64
Click Events	10.50	9.50	4.01

Table 1: Summary of User Interaction Metrics logged by the system. In this experiment, the "log interval setting" setting for mouse movement is set to 100. s= seconds, px = pixels.

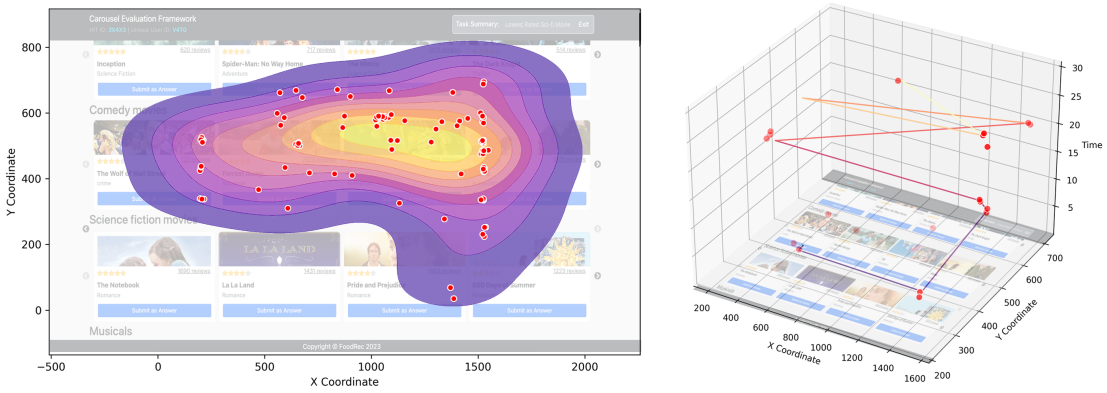

Figure 3: Multifaceted Visualization of User Interactions. Left: a two-dimensional density map of click events across the interface for all participants, highlighting areas of concentrated activity. Right: a 3D trajectory of mouse movements over time for one random participant, offering a dynamic perspective on user engagement with the task interface.

and locations. Red dots represent actual click events, highlighting user activity areas.

The right panel in Figure 3 presents a three-dimensional visualization of mouse movement data from one user, showing the spatial and temporal dynamics of interactions. It reveals user engagement patterns, usability bottlenecks, and intuitive design elements. In CARE , mouse movement serves as a proxy for user engagement. While not as precise as eye tracking, it provides substantial interaction information, especially in well-designed experiments.

5 DISCUSSION

In this paper, we discuss the development of CARE , a new infrastructure for evaluating carousel-based interfaces in recommender systems. Our motivation originated from a gap we identified in the existing literature on standardized analysis tools for these interfaces. Our preliminary analysis of log data from a proof-of-concept user experiment revealed interesting patterns of user interaction and behavior.

Our recognition of the limitations of our current setup has led to plans for further expansion. With an open-source approach [1], we invite collaboration and enhancements from the research community. Furthermore, we share the data analysis code used in this study, inviting others to contribute their findings and methodologies too.

One limitation of our work is, that despite efforts to optimize log collection and configurable settings, we advise limiting experiments to under five minutes per task. This manages data volume effectively, as longer tasks generate substantial logs, potentially impacting system performance and analysis.

We also recognize that CARE system currently does not support many different types of usability testing scenarios and is limited to predefined task completion. We plan to add more options to the system as our research progresses. We also hope that other researchers contribute to this project by adding more types of studies.

Finally, we use mouse movement as a proxy for user interaction, which has limitations like lack of mobile device support and discrepancies between intention and movement. Proper experiment design can mitigate some issues. We plan to include "WebGazer" [12], a browser-based eye tracker, in CARE , enhancing compatibility, particularly for mobile devices. Our ambition with this project is to lower the barriers to entry for research in this field, making these tools available to a wider research audience, and establishing a standardized testing environment. This environment aims to support replicable and comparable studies, ultimately contributing to a more coherent body of research in carousel interface usability.

[1] https://github.com/benrahdari/carousel-eval

REFERENCES

[1] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso

Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 2405–2414. https://doi.org/10.1145/3404835.3463245

[2] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. V-Elliot: Design, Evaluate and Tune Visual Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 768–771. https://doi.org/10.1145/3460231.3478881

[3] Federico Bianchi, Patrick John Chia, Jacopo Tagliabue, Ciro Greco, Gabriel S P Moreira, Davide Eynard, Fahd Husain, and Claudio Pomo. 2023. EvalRS 2023: Well-Rounded Recommender Systems for Real-World Deployments. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 5851–5852. https://doi.org/10.1145/3580305.3599222

[4] Robin Burke, Michael Ekstrand, Daniel Kluver, Bart Knijnenburg, Joseph Konstan, and Edward Malthouse. 2024. POPROX: Platform for OPen Recommendation and Online eXperimentation. https://poprox.ai/

[5] Nicolas Burny. 2020. Towards Supporting Reproducibility of Experimental Studies in GUI Visual Design. In *Companion Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Sophia Antipolis, France) *(EICS '20 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 13, 4 pages. https://doi.org/10.1145/3393672.3398644

[6] Patrik Dokoupil and Ladislav Peska. 2023. EasyStudy: Framework for Easy Deployment of User Studies on Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1196–1199. https://doi.org/10.1145/3604915.3610640

[7] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. Measuring the User Satisfaction in a Recommendation Interface with Multiple Carousels. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 212–217. https://doi.org/10.1145/3452918.3465493

[8] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 257–260. https://doi.org/10.1145/1864708.1864761

[9] Dietmar Jannach, Michael Jugovac, and Lucas Lerche. 2012. Recommender systems beyond accuracy: A multi-criteria evaluation approach. In *Recommender Systems Handbook*. Springer Berlin Heidelberg, Berlin, Germany, 297–331.

[10] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[11] Benedikt Loepp and Jürgen Ziegler. 2023. How Users Ride the Carousel: Exploring the Design of Multi-List Recommender Interfaces From a User Perspective. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1090–1095. https://doi.org/10.1145/3604915.3610638

[12] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, ACM, New York, NY, 3839–3845. https://doi.org/10.5555/3061053.3061156

[13] Behnam Rahdari, Peter Brusilovsky, Daqing He, Khushboo Thaker, Zhimeng Luo, and Young Ji Lee. 2022. Helper: an interactive recommender system for ovarian cancer patients and caregivers. In *16th ACM Conference on Recommender Systems*. ACM, New York, NY, 644–647. https://doi.org/10.1145/3523227.3551471

[14] Behnam Rahdari, Peter Brusilovsky, and Branislav Kveton. 2024. Towards Simulation-Based Evaluation of Recommender Systems with Carousel Interfaces. *ACM Trans. Recomm. Syst.* 1, 1 (jan 2024), 1. https://doi.org/10.1145/3643709

[15] Behnam Rahdari, Peter Brusilovsky, and Alireza Javadian Sabet. 2021. Controlling Personalized Recommendations in Two Dimensions with a Carousel-Based Interface. In *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS'21) at 2021 ACM Conference on Recommender Systems (RecSys'21) (CEUR Workshop Proceeding, Vol. 2948)*. CEUR, Online, 112–122.

[16] Behnam Rahdari, Branislav Kveton, and Peter Brusilovsky. 2022. From Ranked Lists to Carousels: A Carousel Click Model. *arXiv preprint arXiv:2209.13426* 1 (2022), 1.

[17] 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*. ACM, New York, YN, 166–174. https://doi.org/10.1145/3511095.3531278

[18] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *J. Mach. Learn. Res.* 21, 1, Article 95 (jan 2020), 5 pages.

[19] Tobias Schnabel, Paul N. Bennett, Susan T. Dumais, and Thorsten Joachims. 2018. Short-Term Satisfaction and Long-Term Coverage: Understanding How Users Tolerate Algorithmic Exploration. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 513–521. https://doi.org/10.1145/3159652.3159700