FINN.no Slates Dataset: A new Sequential Dataset Logging Interactions, all Viewed Items and Click Responses/No-Click for Recommender Systems Research

Simen Eide
simen.eide@finn.no
University of Oslo and FINN.no
Oslo, Norway

Arnoldo Frigessi
arnoldo.frigessi@medisin.uio.no
University of Oslo
Oslo, Norway

Helge Jenssen
helge.jenssen@finn.no
FINN.no
Oslo, Norway

David S. Leslie
d.leslie@lancaster.ac.uk
University of Lancaster
Lancaster, United Kingdom

Joakim Rishaug
joakim.rishaug@finn.no
FINN.no
Oslo, Norway

Sofie Verrewaere
Sofie.verrewaere@schibsted.com
Schibsted ASA
Oslo, Norway

CCS CONCEPTS

• **Information systems → Personalization**.

KEYWORDS

slate recommendations, search result, candidate sampling, marketplace data, reinforcement learning, bandit, item attributes, off-policy

1 INTRODUCTION

We present a novel recommender systems dataset that records the sequential interactions between users and an online marketplace. The users are sequentially presented with both recommendations and search results in the form of ranked lists of items, called slates, from the marketplace (figure 1). The dataset includes the presented slates at each round, whether the user clicked on any of these items and which item the user clicked on. Although the usage of exposure data in recommender systems is growing [2, 5, 6, 9], to our knowledge there is no open large-scale recommender systems dataset that includes the slates of items presented to the users at each interaction. As a result, most articles on recommender systems do not utilize this exposure information. Instead, the proposed models only depend on the user's click responses, and assume that the user is exposed to all the items in the item universe at each step, often called uniform candidate sampling ([3, 7, 8]). This is an incomplete assumption, as it takes into account items the user might not have been exposed to. This way items might be incorrectly considered as

not of interest to the user. Taking into account the actually shown slates allows the models to use a more natural likelihood, based on the click probability given the exposure set of items, as is prevalent in the bandit and reinforcement learning literature [1, 10, 13, 14, 17]. [6] shows that likelihoods based on uniform candidate sampling (and similar assumptions) are implicitly assuming that the platform only shows the most relevant items to the user. This causes the recommender system to implicitly reinforce feedback loops and to be biased towards previously exposed items to the user.

We hope that the release of this dataset can assist researchers to construct more realistic user preference models by utilizing the exposed slates and contribute to more robust offline evaluation criteria using for example importance sampling [12]. The dataset along with quickstarts and tutorials are available at https://github.com/finn-no/recsys-slates-dataset.
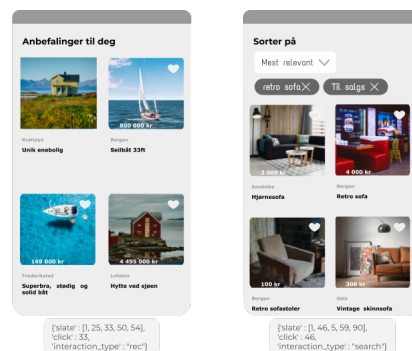
Figure 1: An example screenshot of the logged slates as they appear on FINN.no. Left shows a recommendation slate. Right shows the resulting slate of a search query. In both cases four items are exposed to the user. We also provide examples (below) of how these slates would have been logged in the dataset: "slate" is a list of the viewed items including the special item of "no click" (which is the index 1), "click" is the index of the clicked item, and "interaction_type" encodes whether it is a search or recommendation slate.

1.1 Related datasets

There are some datasets on ad display or information retrieval [4, 15]. However, these datasets have some shortcomings. Failing to identify the user or lacking a sufficient number of interactions per user cause them to be unfitting for benchmark analyses. Parallel to our work, [16] released a dataset containing sessions of user interactions with an eCommerce site. It includes a rich set of information using featureized query, description and image information. The dataset reports the topK search results in each interaction, working as an approximation to the presented slates that we report in this dataset. There are also simulation environments that allow interaction with synthetic users [11], showing the need for real-world datasets.

1.2 About the marketplace

FINN.no is the leading marketplace in Norway for real estate, job offerings, vehicles and general merchandise. It is also one of the largest websites in the country. Approximately 20% of the clicks on item listings come from its recommender systems, and the system recommends around 90 million items each day. FINN.no is part of the media group Schibsted ASA.

1.3 Detailed description of the dataset

The present dataset was collected over 30 days on the marketplace FINN.no. The dataset contains 37.4 million logged search or recommendation user interactions. For each item, we log all the items presented in the slate and the user response. In addition, we provide a text string attribute for each item containing its categories. Approximately 70% of the interactions originate from search queries, the remaining 30% from recommendations. 24.4% of the interactions resulted in the user not clicking on any item. The dataset consists of 2.3 million users and 1.3 million unique items. We provide more dataset statistics in Table 1.

Interaction data. The dataset contains the interaction types, the presented slates and the user responses. The interaction type identifies whether it was a search query or the recommendation system that presented the slate. The presented slate contains all items that were shown to the user during each interaction (the full scroll length by the user), but not necessarily in the observed order. For example, in figure 1, we would have logged 4 items in both cases. Finally, the dataset also includes the user response which may be one of the items in the presented slate or the action of not clicking on any of the items (*no-click*).

Item attributes. The items belong to 290 unique groups that are constructed using a combination of categorical information and the geographical location describing where the item is made available in the marketplace. Each group label describes this information using a text string. For example, the group *BAP, antiques, Trøndelag* is interpreted as belonging to the general merchandise ("Bits And Pieces") category, in the antique category and being sold in the county of *Trøndelag*. Similarly, the group *MOTOR, Sogn og Fjordane* is interpreted as the item belonging to the motor category, and sold in the county *Sogn og Fjordane*. Note that in this group the categorical information is omitted, to keep the groups sufficiently large for anonymization requirements.

Table 1: Dataset statistics

Description	Value
Number of interactions	37.5m
Length of time period of data collection	30 days
Unique Items	1.3m
Unique item groups	290
Unique Users	2.3m
Average number of interactions per user	16.4
Rate of interactions with *no-click*	24.4%
Average Slate length	10.1 items
Percentage of search produced slates	69.7%
Percentage of recommendation produced slates	30.3%

Processing and anonymization. The published dataset has been processed to anonymize its content and to enhance computation. Due to the abundant amount of *no-click* signals (which were 76.3% of the total number of interactions), the *no-click* interactions were first uniformly down-sampled to 10% of the original number of interactions. Users with less than 10 interactions during the data collection period have been removed. If a user had more than 20 interactions, only the 20 first occurring interactions are logged. The maximum number of items considered in a slate is capped at 25 items. This capping impacts only 5% of the slates. If the user clicked on any of the items that were removed, we removed the whole interaction from the dataset. These simplifications only affect a small portion of the dataset and do not, in our view, alter the main contribution of the dataset.

Train/Validation/Test Split. We propose a way to split the dataset into training, validation and test dataset. This is also the split used to evaluate the models in [6]. Let all interactions for 90% of the users be placed in the training dataset. For the remaining 10% of the users, we place 5% in each validation and test dataset. However, the first five interactions per user are placed in the training dataset. This allows a recommender system to estimate some user-specific parameters based on limited knowledge of the user. These splits are implemented as *Pytorch DataLoaders* in the repository for easy application.

Limitations. We have explicitly not included some data fields that may have been relevant for recommender systems research, mostly due to privacy concerns. For example, including a full description of items, the text queries that were used in search slates and exact timestamps of all interactions have been excluded because we did not find a way to share this information without the risk of exposing private information. We also limited the total number of interactions for the same reason.

ACKNOWLEDGMENTS

research is also part of the strategic partnership between BigInsight at the University of Oslo and the STORi Centre for Doctoral Training at Lancaster University.

2 SPEAKER BIO

Simen Eide is a research scientist working on recommender systems at the online marketplace FINN.no, where he is also pursuing a PhD degree at the University of Oslo, Norway. His research interests are recommender systems, statistical learning, bayesian statistics, decision making and bandit theory. Eide has worked on building real world recommender systems since 2015. Previously he worked on quantitative risk management, designing portfolio risk models on a variety of asset classes. He holds an MSc in mathematical finance from the University of Oslo.

REFERENCES

[1] Richard S. Sutton Barto and Andrew G. 2015. Reinforcement Learning: An Introduction.
[2] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed Chi. 2018. Top-K Off-Policy Correction for a REINFORCE Recommender System. (2018). https://doi.org/10.1145/3289600.3290999
[3] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16* (2016), 191–198. https://doi.org/10.1145/2959100.2959190
[4] Criteo. 2020. Criteo 1TB Click Logs dataset. https://ailab.criteo.com/download-criteo-1tb-click-logs-dataset/
[5] James Edwards and David Leslie. 2018. Diversity as a Response to User Preference Uncertainty. In *Statistical Data Science*. WORLD SCIENTIFIC (EUROPE), 55–68. https://doi.org/10.1142/9781786345400{_}0004
[6] Simen Eide, David S. Leslie, and Arnoldo Frigessi. 2021. Dynamic Slate Recommendation with Gated Recurrent Units and Thompson Sampling. (2021), 1–30. https://doi.org/10.21203/rs.3.rs-525958/v1
[7] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*. 241–248. https://doi.org/10.1145/2959100.2959167
[8] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 263–272. https://doi.org/10.1109/ICDM.2008.22
[9] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng Tze Cheng, Tushar Chandra, and Craig Boutilier. 2019. SLateq: A tractable decomposition for reinforcement learning with recommendation sets. *IJCAI International Joint Conference on Artificial Intelligence* 2019-Augus (2019), 2592–2599. https://doi.org/10.24963/ijcai.2019/360
[10] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, Jim McFadden, Tushar Chandra, and Craig Boutilier. 2019. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv* (5 2019). http://arxiv.org/abs/1905.12767
[11] Eugene Iey, Chih Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekarx, Jing Wang, Rui Wu, and Craig Boutilier. 2019. RECSIM: A configurable simulation platform for recommender systems. *arXiv* (2019), 1–23.
[12] Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. 2018. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*. http://www.joachims.org/banditnet/
[13] Tor Lattimore and Csaba Szepesvári. 2019. *Bandit Algorithms*. Technical Report.
[14] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra Spotify. 2018. Explore, Exploit, and Ex-plain: Personalizing Explainable Recommendations with Bandits. (2018). https://doi.org/10.1145/3240323.3240354
[15] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. *TripClick: The Log Files of a Large Health Web Search Engine*. Vol. 1. Association for Computing Machinery. http://arxiv.org/abs/2103.07901
[16] Jacopo Tagliabue, Ciro Greco, Jean-Francis Roy, Bingqing Yu, Patrick John Chia, Federico Bianchi, and Giovanni Cassani. 2021. *SIGIR 2021 E-Commerce Workshop Data Challenge*. Vol. 1. Association for Computing Machinery. http://arxiv.org/abs/2104.09423
[17] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Dawei Yin, Yihong Zhao, and Jiliang Tang. 2018. Deep Reinforcement Learning for List-wise Recommendations. 9 (2018). https://doi.org/10.1145/nnnnnnn.nnnnnnn