

Unbiased Learning to Ranking Evaluation Task (ULTRE)

Jiaxin Mao, Qingyao Ai, Junqi Zhang, Tao
Yang, Yurou Zhao, Yiqun Liu

2021.03.29





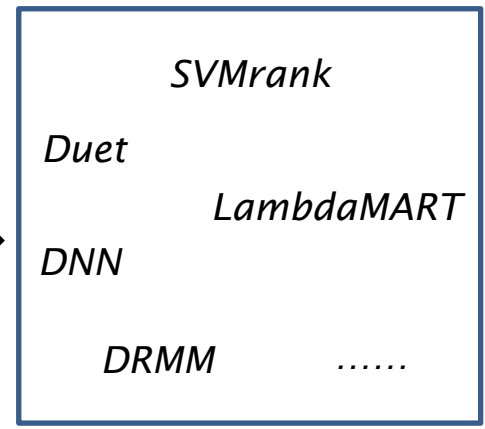
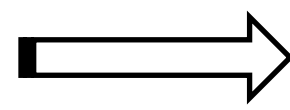
Background



- Unbiased Learning to Rank (ULTR):
 - Users' interaction with search systems can reflect their implicit relevance feedback for search results
 - But they are also noisy and biased
 - ULTR: Learning an unbiased ranker from biased user feedback

User clicks

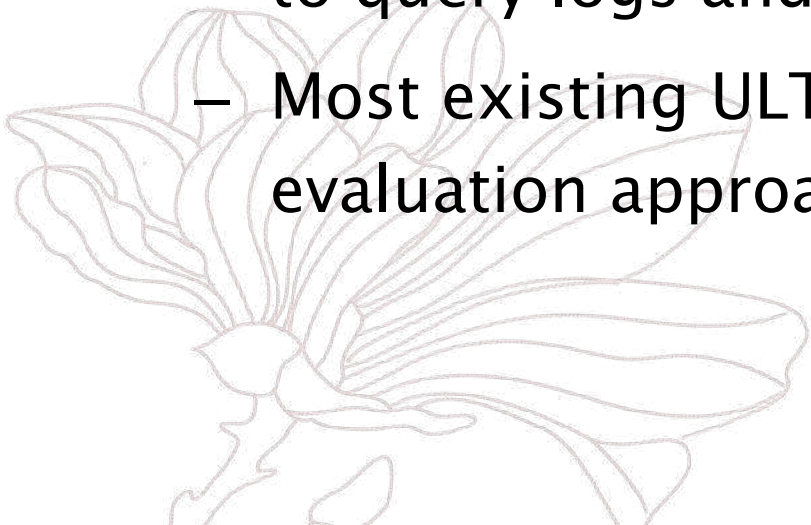
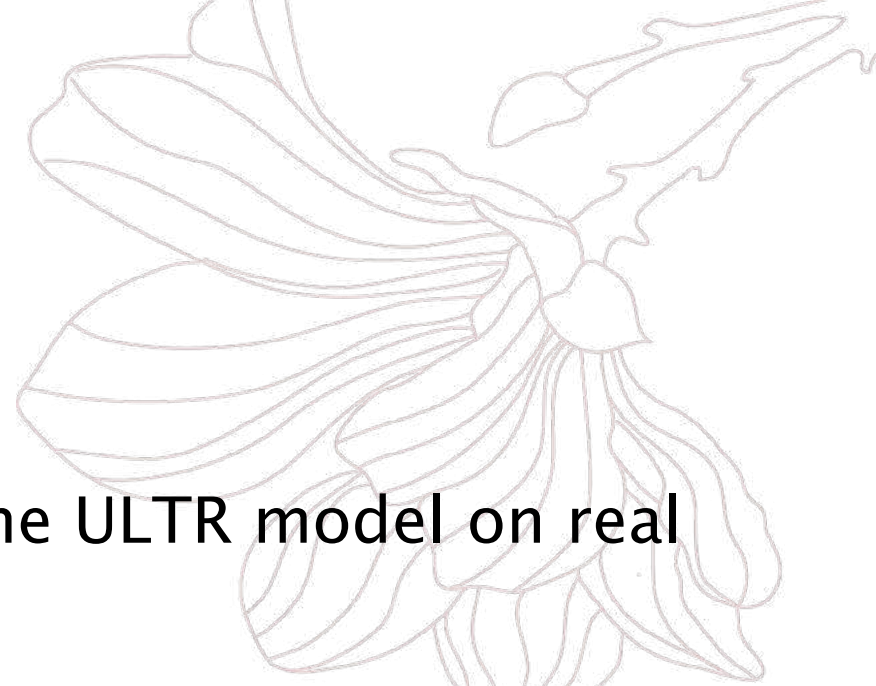
Ranking models





Background

- The evaluation of ULTR
 - Ideally, we should train and evaluate the ULTR model on real search logs and online search systems
 - Not possible for academic researchers due to a lack of access to query logs and online systems
 - Most existing ULTR studies utilize a simulation-based evaluation approach

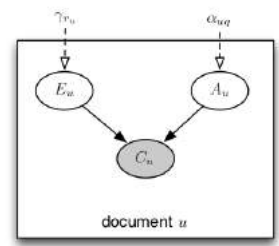




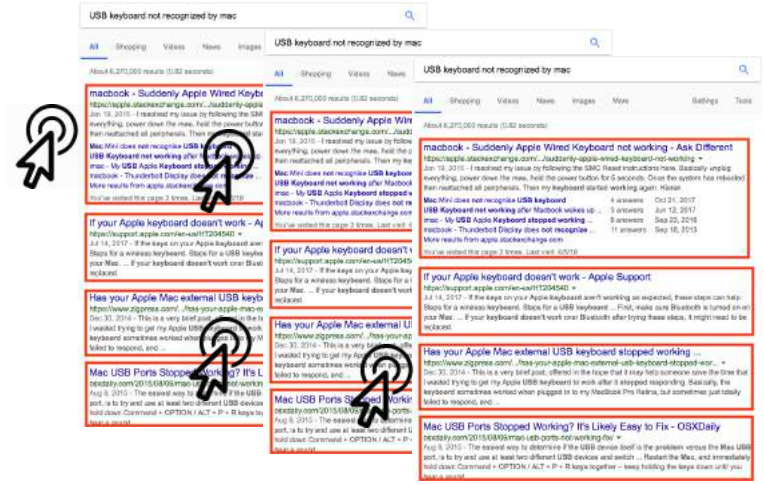
Background

• Simulation-based evaluation of ULTR

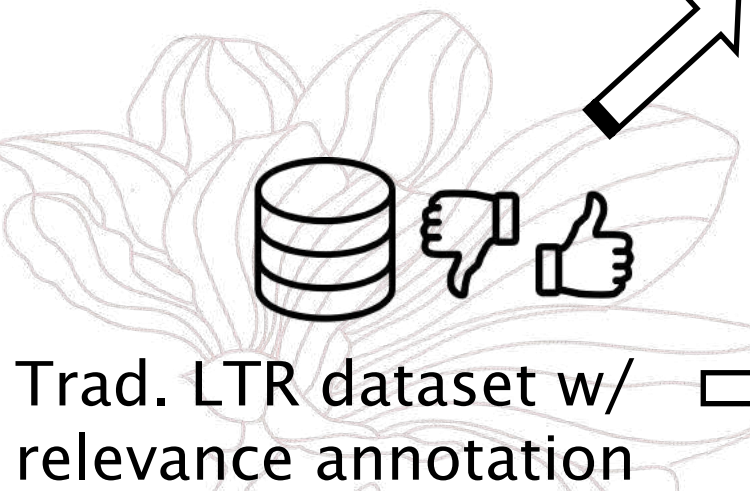
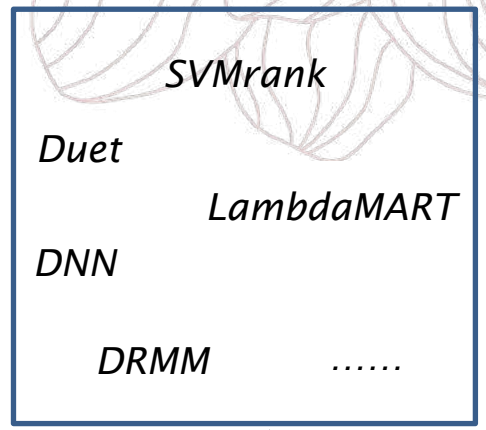
User behavior model



Simulated clicks

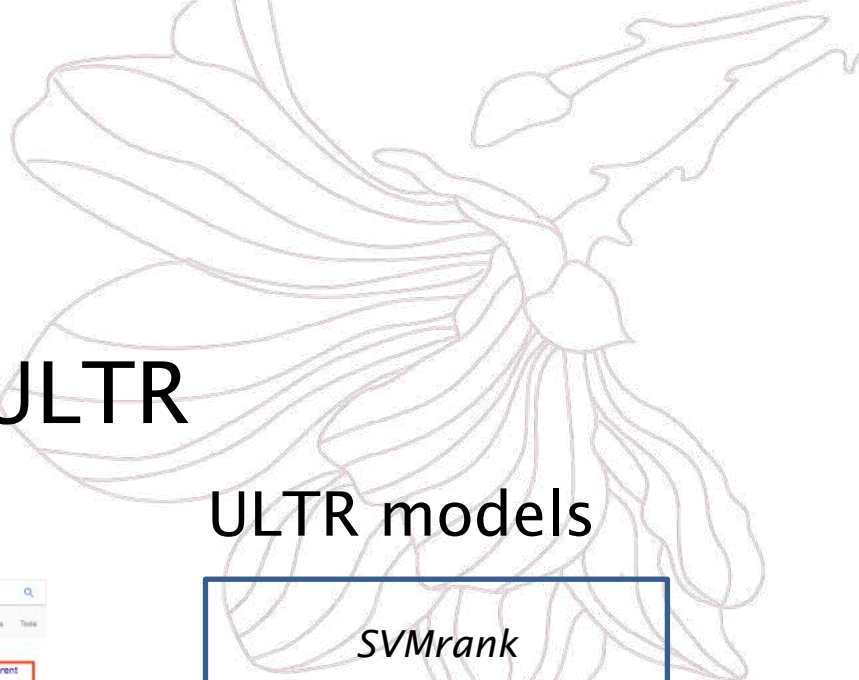


ULTR models



Trad. LTR dataset w/
relevance annotation

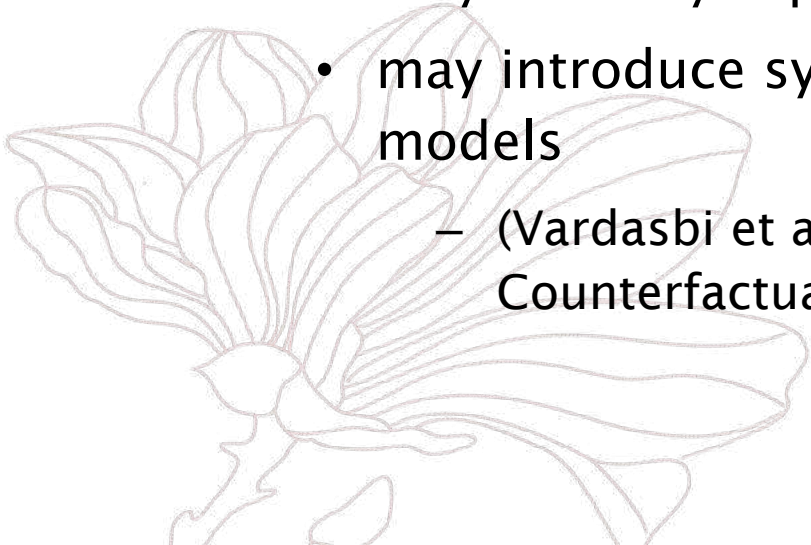
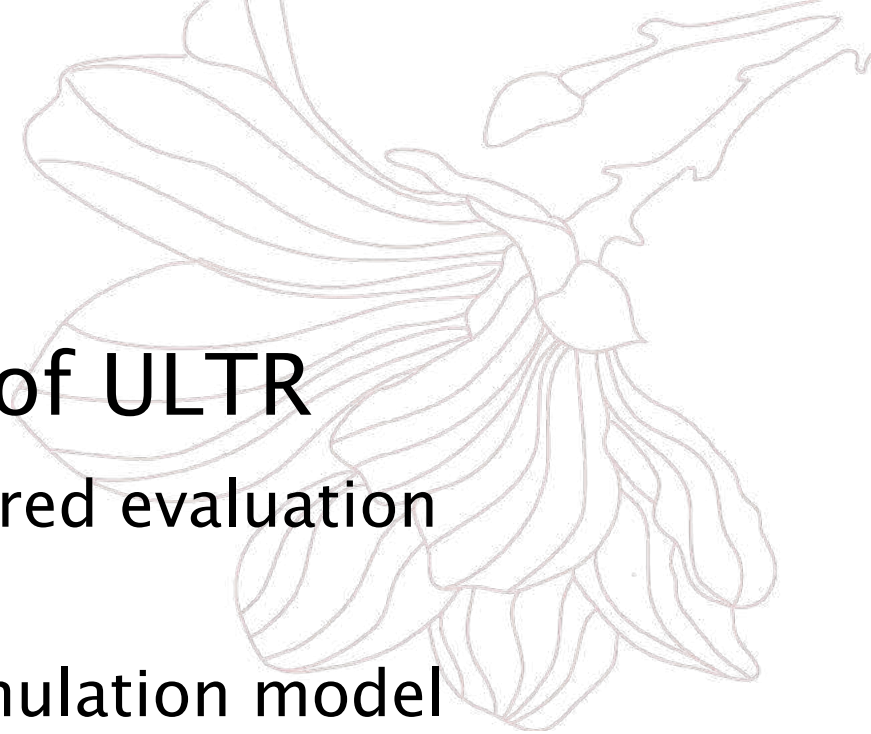
Evaluate





Background

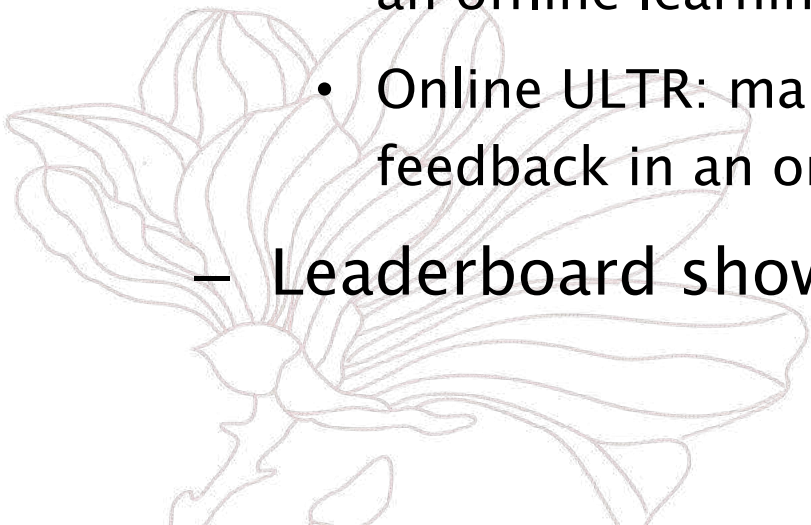
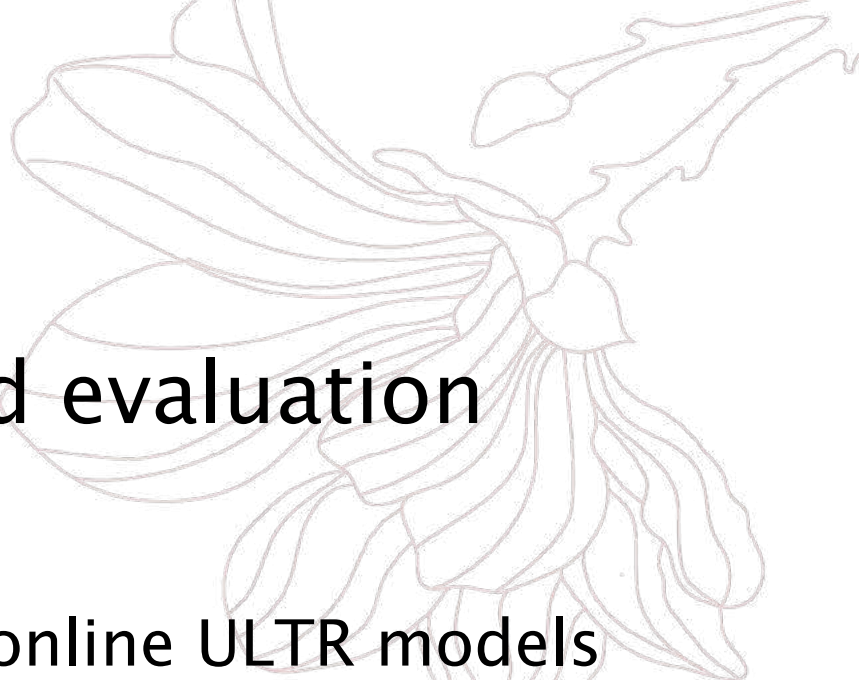
- Limitations with the evaluation of ULTR
 - No standard evaluation settings or shared evaluation benchmarks for the ULTR community
 - Most studies only use a single user simulation model
 - may not fully capture the diverse patterns of real user behavior
 - may introduce systematic biases into the comparison among ULTR models
 - (Vardasbi et al. Cascade Model-based Propensity Estimation for Counterfactual Learning to Rank, SIGIR 2020)





Motivation

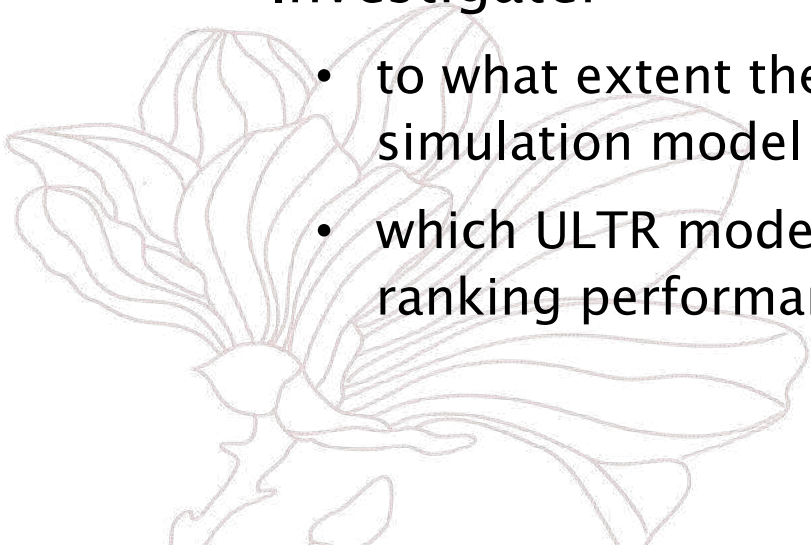
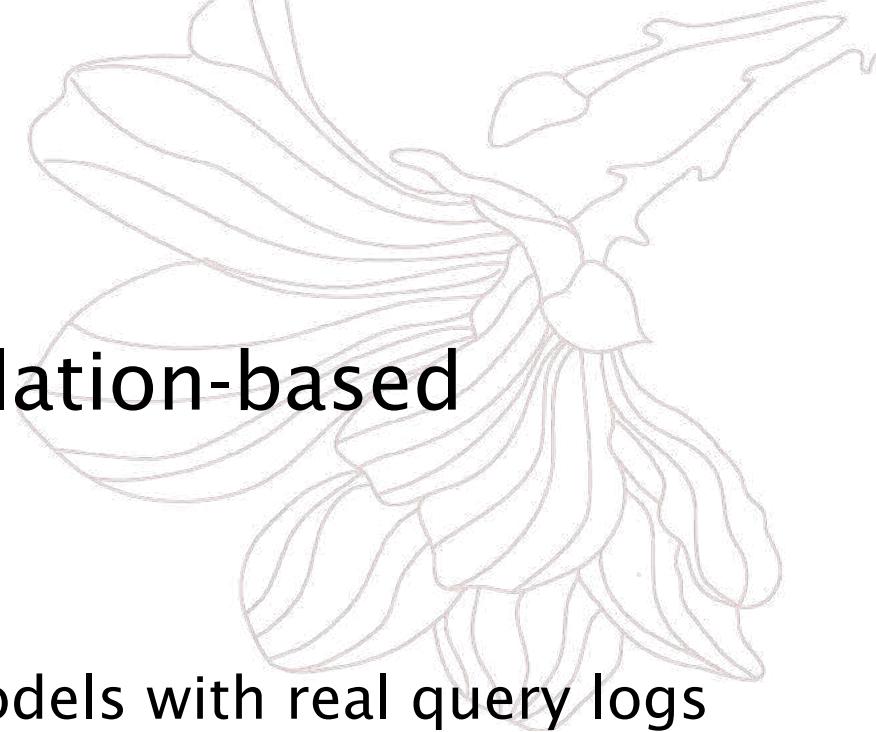
- Provide a shared benchmark and evaluation service for ULTR
 - Support the evaluation of both offline/online ULTR models
 - Offline ULTR: learn an unbiased ranker with biased historical logs in an offline learning manner
 - Online ULTR: make online interventions of ranking and elicit unbiased feedback in an online learning process
 - Leaderboard showing real-time results on the validation set





Motivation

- Investigate and improve the simulation-based evaluation approach
 - Use multiple user simulation models
 - Train and calibrate the user simulation models with real query logs
 - Investigate:
 - to what extent the evaluation results will be influenced by the user simulation model
 - which ULTR model can adapt to different environments and has a robust ranking performance

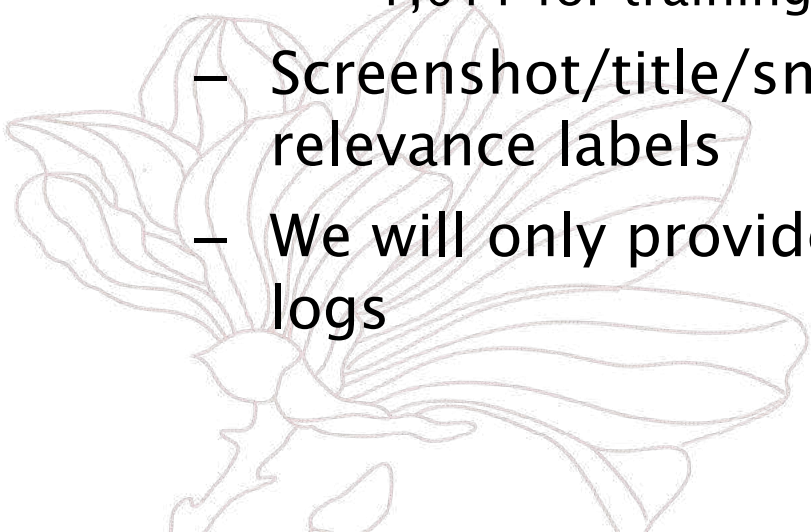




Methodology

- Datasets:

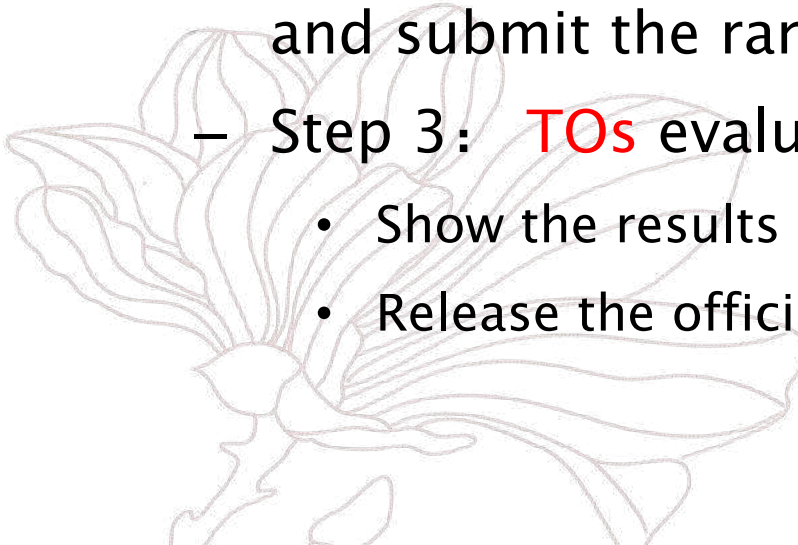
- Based on the Sogou-SRR, a public dataset for relevance estimation and ranking in Web search.
 - <http://www.thuir.cn/data-srr/>
- Select 1,211 queries with at least 10 successfully crawled results
 - 1,011 for training, 100 for validation, 100 for testing
- Screenshot/title/snippet/HTML/parsed HTML tree/4-level human relevance labels
- We will only provide the simulated clicks, *not* the genuine click logs





Methodology

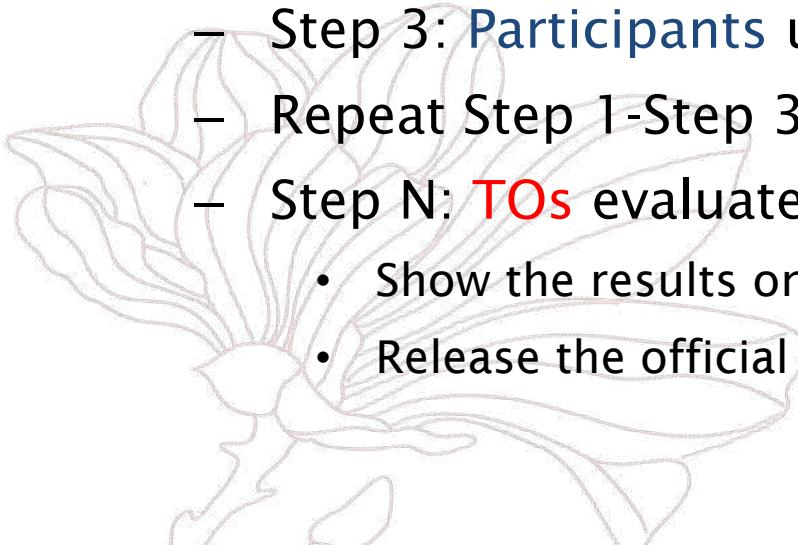


- Evaluation protocol for offline UTLR models:
 - Step 1: **TOs** generate simulated click logs for all training queries
 - Use different click models (PBM/DCM/UBM/MCM/...)
 - Train and calibrate the click models with real click logs
 - Step 2: **Participants** train UTLR models with simulated click logs and submit the ranking lists (runs) for validation/test queries
 - Step 3: **TOs** evaluate the runs
 - Show the results on validation set on the leaderboard
 - Release the official results on test set in the final report
- 



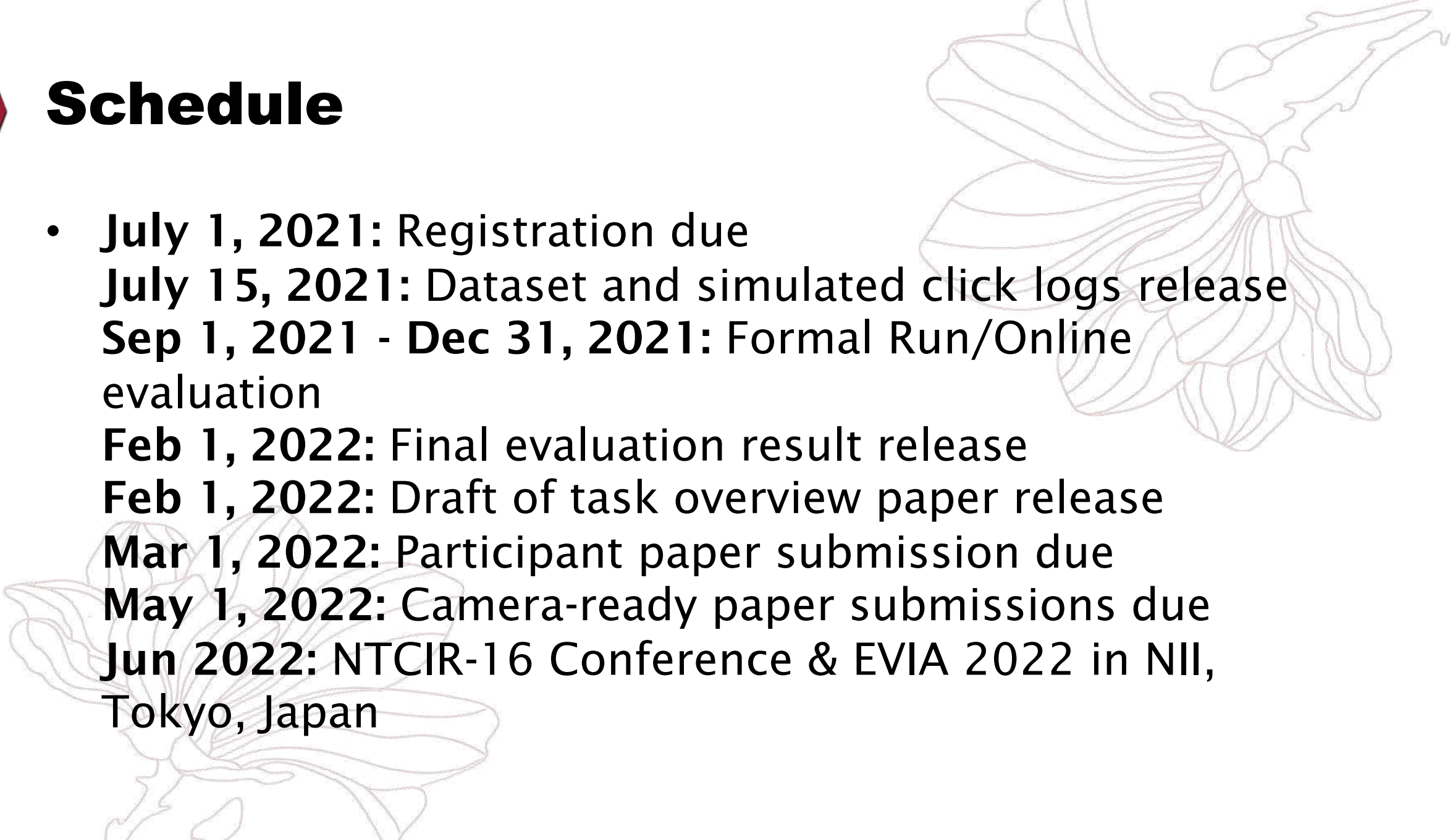
Methodology



- Evaluation protocol for online UTLR models:
 - Step 1: **Participants** submit the ranking lists for training/validation/test queries
 - And specify that they want to receive x% of impressions
 - Step 2: **TOs** sample training queries and generate simulated clicks on the ranking lists submitted by participants
 - Step 3: **Participants** update their models with the simulated clicks
 - Repeat Step 1-Step 3 until **participants** receive 100% of impressions
 - Step N: **TOs** evaluate results on validation/test set
 - Show the results on validation set on the leaderboard
 - Release the official results on test set in the final report
- 



Schedule

- **July 1, 2021:** Registration due
 - July 15, 2021:** Dataset and simulated click logs release
 - Sep 1, 2021 - Dec 31, 2021:** Formal Run/Online evaluation
 - Feb 1, 2022:** Final evaluation result release
 - Feb 1, 2022:** Draft of task overview paper release
 - Mar 1, 2022:** Participant paper submission due
 - May 1, 2022:** Camera-ready paper submissions due
 - Jun 2022:** NTCIR-16 Conference & EVIA 2022 in NII, Tokyo, Japan
- 



Thanks

Jiaxin Mao, Qingyao Ai, Junqi Zhang, Tao Yang, Yurou
Zhao, Yiqun Liu

maojiaxin@gmail.com

ultre-org@googlegroups.com

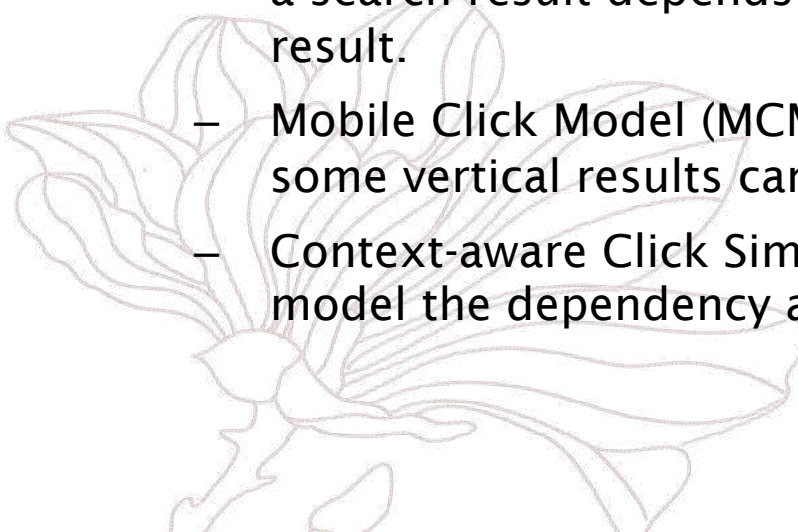
2021.03.29





User Simulation Models



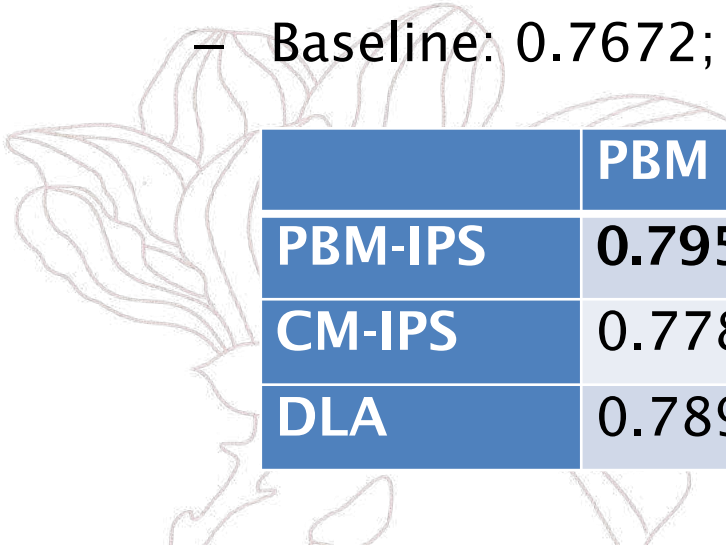
- We plan to use these models in generating simulated clicks:
 - Position-Based Model (PBM): a click model that assumes the click probability of a search result only depends on its relevance and its ranking position.
 - Dependent Click Model (DCM): a click model that is based on the cascade assumption that the user will sequentially examine the results list and find attractive results to click until she feels satisfied with the clicked result.
 - User Browsing Model (UBM): a click model that assumes the examination probability on a search result depends on its ranking position and the distance to the last clicked result.
 - Mobile Click Model (MCM): a click model that considers the click necessity bias (i.e. some vertical results can satisfy users' information need without a click) in user clicks.
 - Context-aware Click Simulator (CCS): a neural click model that uses a two-level RNN to model the dependency among search results and the influence of previous clicks.
- 



Some initial evaluation results



- The ranking performance of three ULTR models:
 - PBM-IPS/ CM-IPS / DLA
- when trained on the click logs generated by:
 - PBM/ DCM/ UBM/ MCM
- Evaluation metric: nDCG@5
 - Baseline: 0.7672; Skyline: 0.8047 (train with rel. labels)



	PBM	DCM	UBM	MCM
PBM-IPS	0.7952	0.7688	0.7792	0.7588
CM-IPS	0.7780	0.7828	0.773	0.7602
DLA	0.7898	0.7800	0.7868	0.7868