ULTRE framework: a framework for Unbiased Learning to Rank Evaluation based on simulation of user behavior

Yurou Zhao[1], Jiaxin Mao[1], Qingyao Ai[2]

[1] Renmin University of China, China
[2] University of Utah, USA

Outline

- Background
  - Unbiased Learning to Rank (ULTR)
  - User simulation-based evaluation approach of evaluating ULTR models and its limitations
- ULTRE (Unbiased Learning to Rank Evaluation) framework
  - two affiliated evaluation protocols
- Experiments
- Conclusion and Future works
- NTCIR16-ULTRE task (Advertisement time)

Background

- Unbiased Learning to Rank (ULTR):
  - Users' interaction with search systems can reflect their implicit relevance feedback for search results
  - Cheap but biased
  - ULTR: Learning an unbiased ranker from biased user feedback
  - Counterfactual/Offline LTR and Online LTR

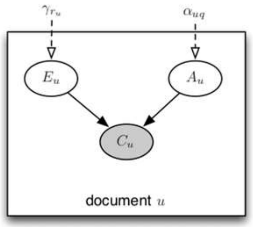
Background

- 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

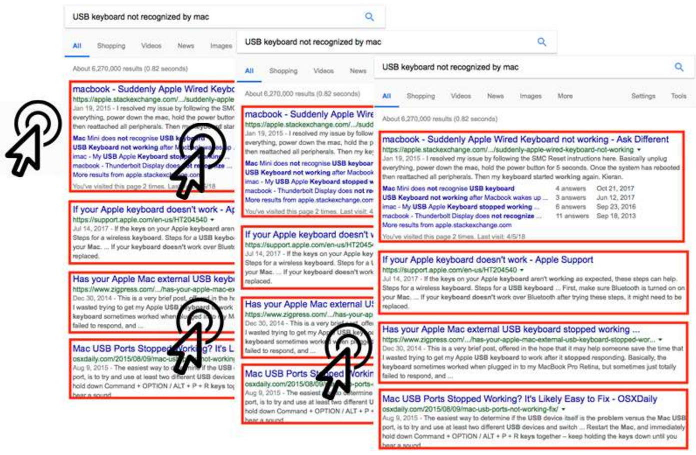
Background

- Simulation-based evaluation of ULTR

User behavior model



Simulated clicks



USB keyboard not recognized by mac

All Shopping Videos News Images More Settings Tools

About 6,270,000 results (0.82 seconds)

macbook - Suddenly Apple Wired Keyboard not working - Ask Different
https://apple.stackexchange.com/.../suddenly-apple-wired-keyboard-not-working
Jan 19, 2015 - I resolved my issue by following the SMC Reset instructions here. Basically unplug everything, power down the mac, hold the power button for 5 seconds. Once the system has rebooted then reattached all peripherals. Then my keyboard started working again. Kieran.

Mac Mini does not recognise USB keyboard	4 answers	Oct 21, 2017
USB Keyboard not working after Macbook wakes up ...	3 answers	Jun 12, 2017
imac - My USB Apple Keyboard stopped working ...	6 answers	Sep 23, 2016
macbook - Thunderbolt Display does not recognize ...	11 answers	Sep 18, 2013

More results from apple.stackexchange.com
You've visited this page 2 times. Last visit: 4/5/18

If your Apple keyboard doesn't work - Apple Support
https://support.apple.com/en-us/HT204540
Jul 14, 2017 - If the keys on your Apple keyboard aren't working as expected, these steps can help. Steps for a wireless keyboard. Steps for a USB keyboard ... First, make sure Bluetooth is turned on on your Mac. ... If your keyboard doesn't work over Bluetooth after trying these steps, it might need to be replaced.

Has your Apple Mac external USB keyboard stopped working ...
https://www.zigpress.com/.../has-your-apple-mac-external-usb-keyboard-stopped-wor...
Dec 30, 2014 - This is a very brief post, offered in the hope that it may help someone save the time that I wasted trying to get my Apple USB keyboard to work after it stopped responding. Basically, the keyboard sometimes worked when plugged in to my MacBook Pro Retina, but sometimes just totally failed to respond, and ...

Mac USB Ports Stopped Working? It's Likely Easy to Fix - OSXDaily
osxdaily.com/2015/08/09/mac-usb-ports-not-working-fix/
Aug 9, 2015 - The easiest way to determine if the USB device itself is the problem versus the Mac USB port, is to try and use at least two different USB devices and switch ... Restart the Mac, and immediately hold down Command + OPTION / ALT + P + R keys together – keep holding the keys down until you hear a sound ...

ULTR models

SVMrank

Duet

LambdaMART

DNN

DRMM ……

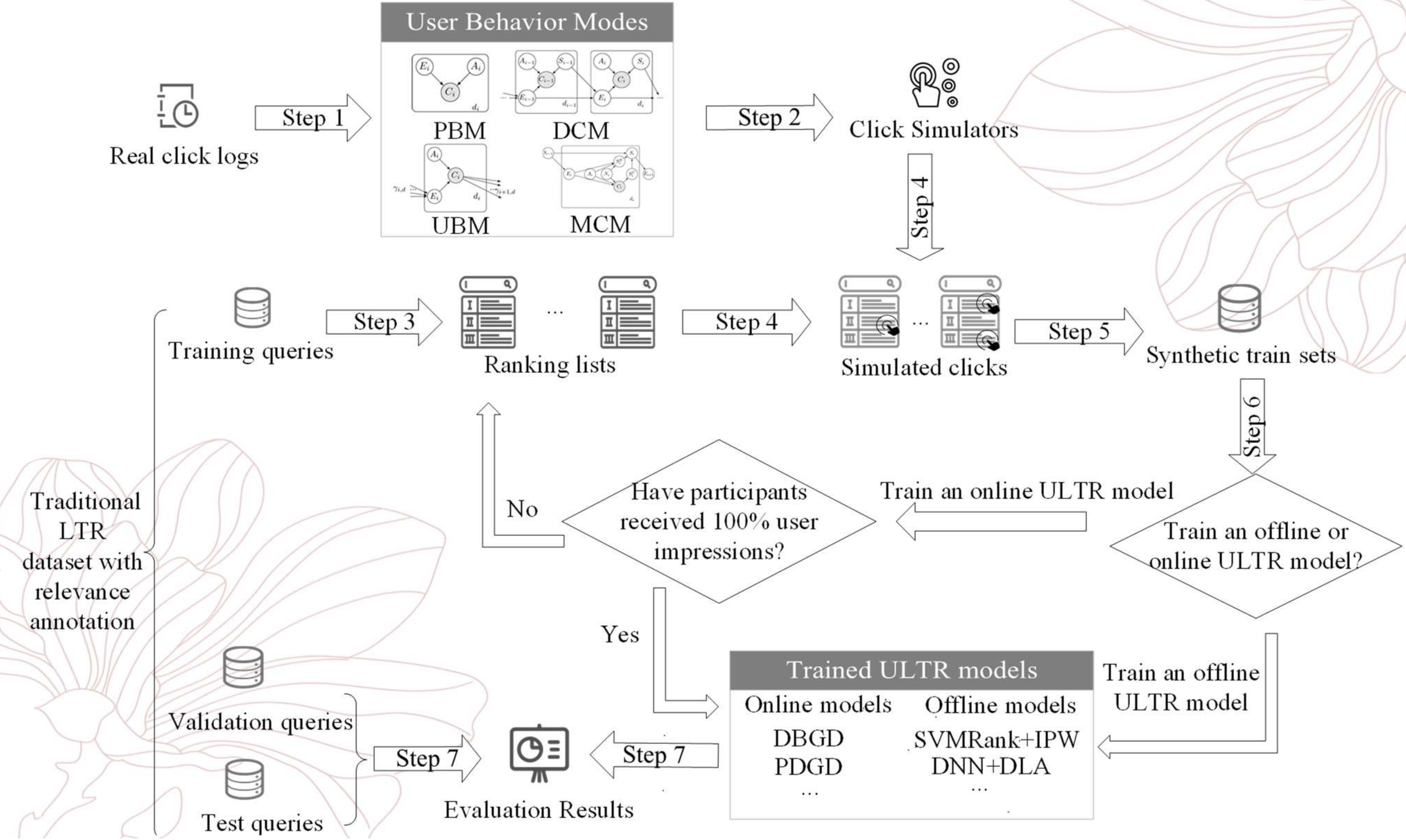
Trad. LTR dataset w/
relevance annotation

Evaluate

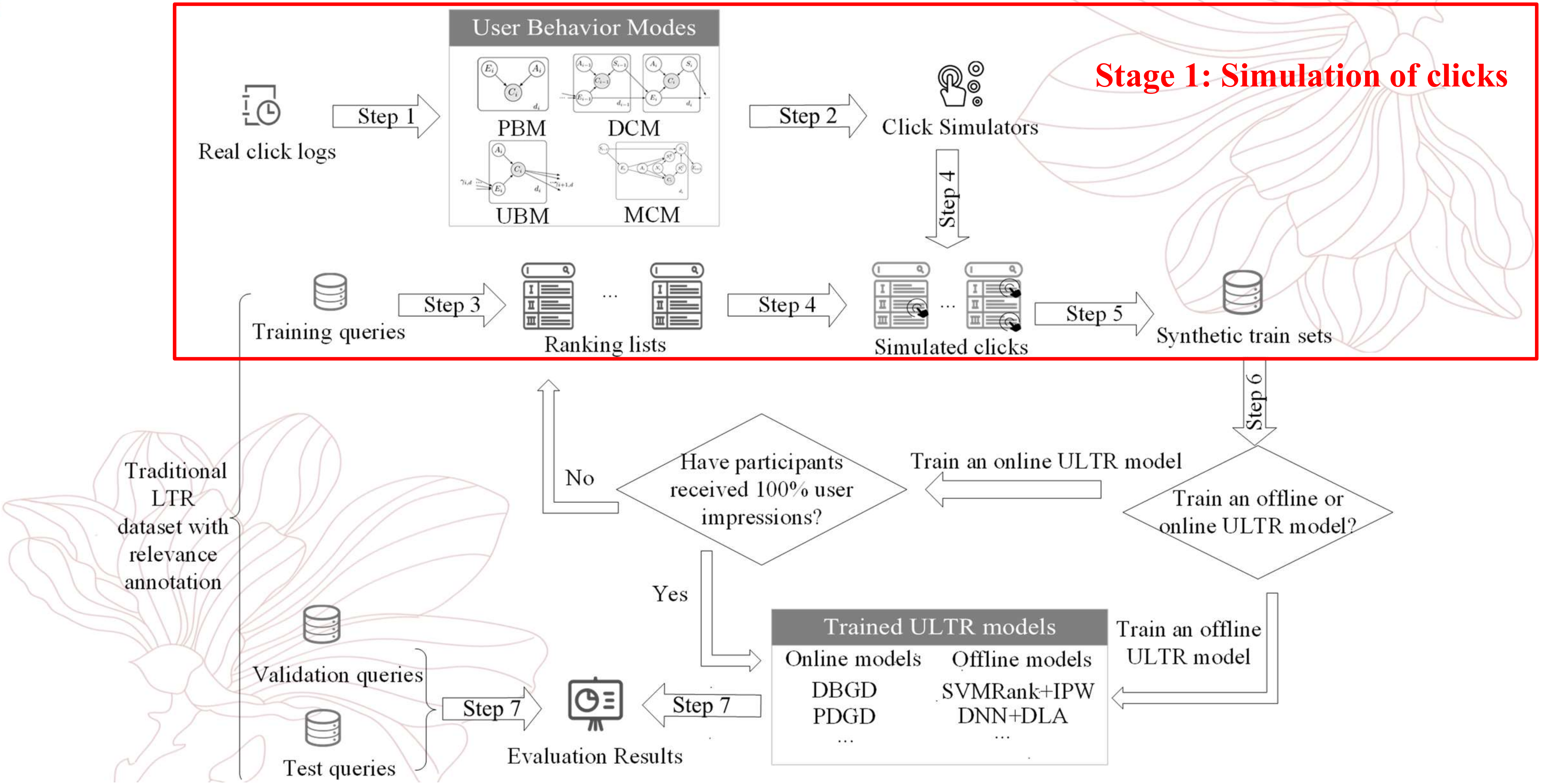
Background

- 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)

ULTRE Framework

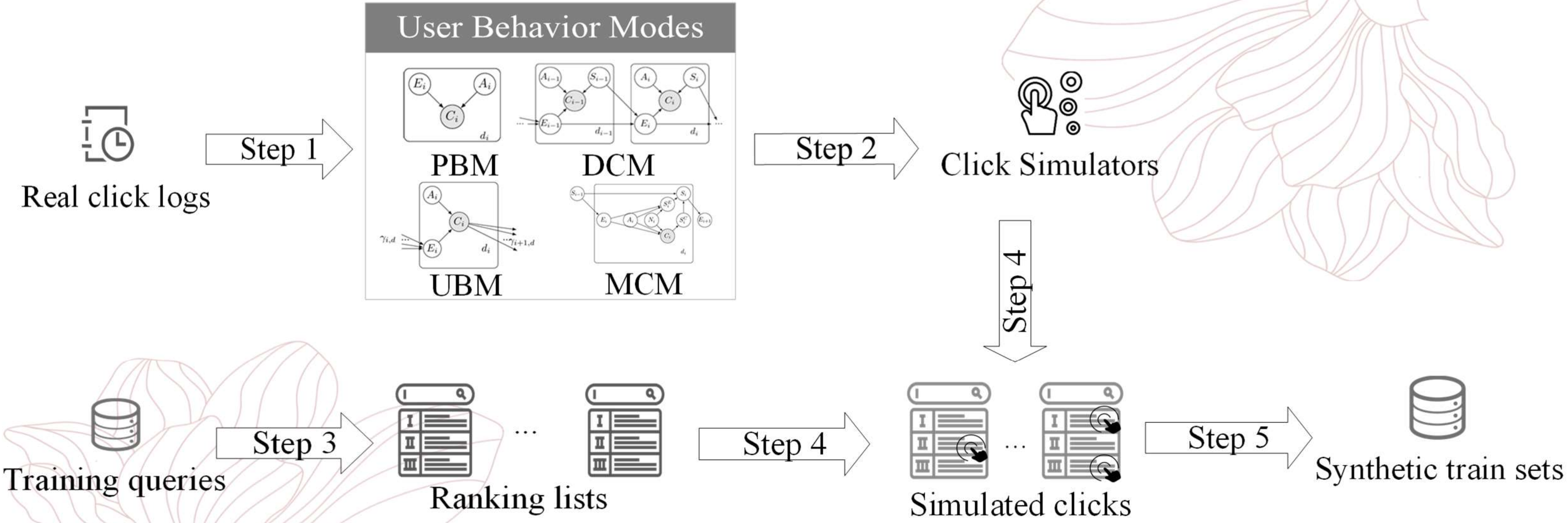


ULTRE Framework - Stage 1



Stage 1: Simulation of clicks

ULTRE Framework - Stage1



User Behavior Modes

PBM

DCM

UBM

MCM

Real click logs

Step 1

Step 2

Click Simulators

Step 4

Training queries

Step 3

Ranking lists

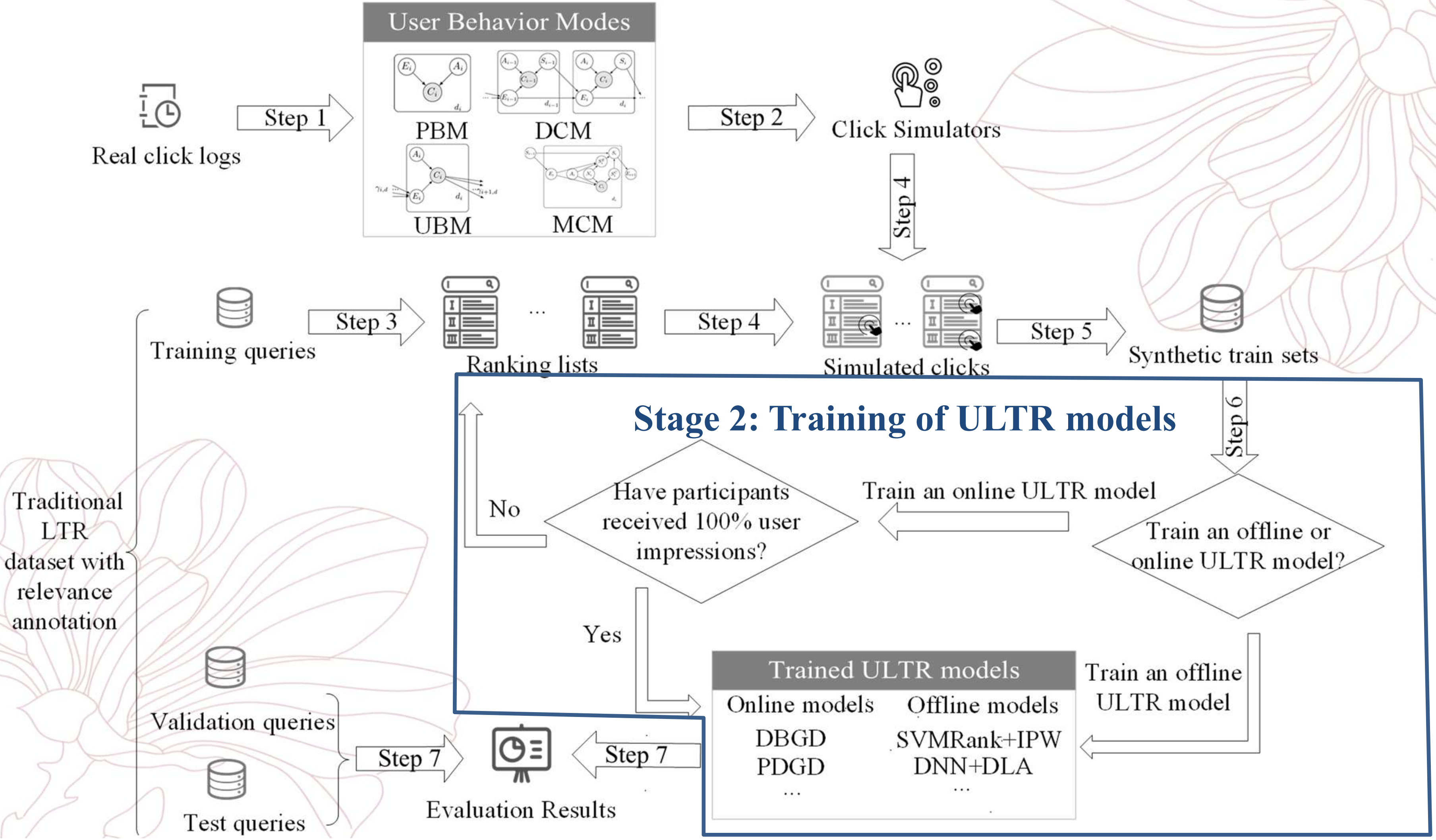
Step 4

Simulated clicks

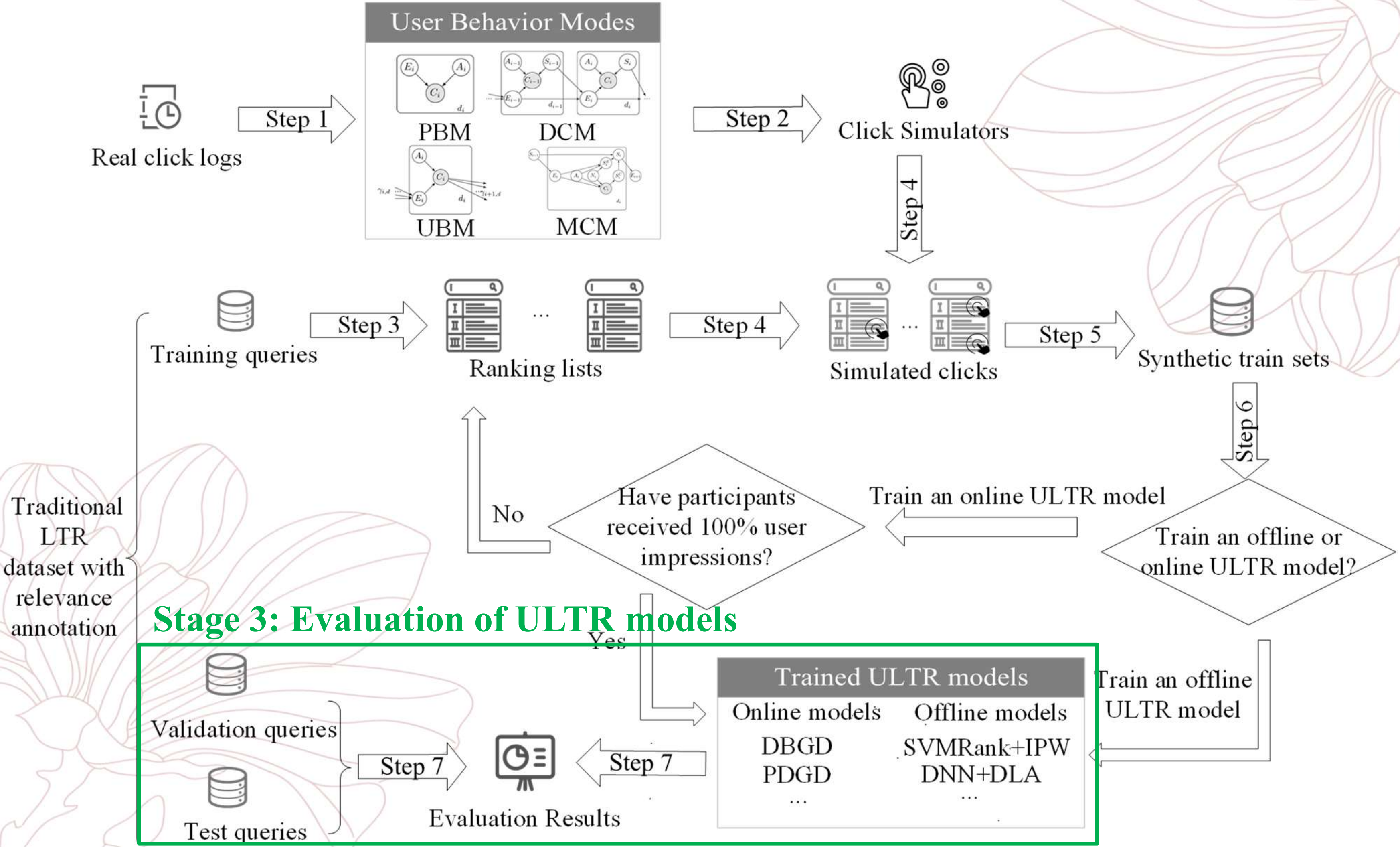
Step 5

Synthetic train sets

ULTRE Framework - Stage 2



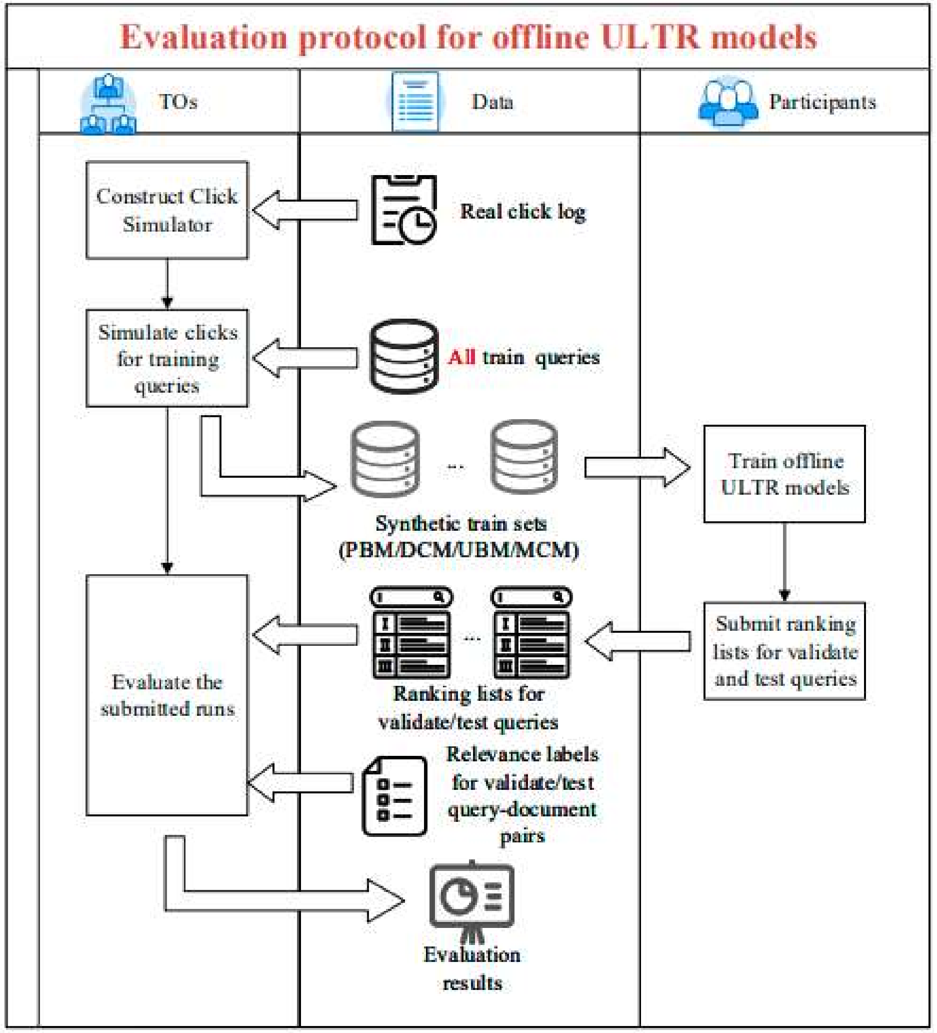
ULTRE Framework - Stage 3



Application of ULTRE framework-- Unbiased Learning to Rank Evaluation Task (ULTRE) in NTCIR-16

- Provide a shared evaluation task and benchmark for the evaluation of different ULTR
- Two evaluation protocols for the ULTRE task
  - Describe how the task organizers (i.e. TOs) interact with the participants and work together to evaluate the ULTR models
  - Offline protocol for offline/counterfactual LTR models
  - Online protocol for online LTR models

Evaluation protocol for offline UTLR models


Evaluation protocol for offline ULTR models
TOs
Data
Participants
Construct Click Simulator
Real click log
Simulate clicks for training queries
All train queries
Synthetic train sets (PBM/DCM/UBM/MCM)
Train offline ULTR models
Submit ranking lists for validate and test queries
Ranking lists for validate/test queries
Evaluate the submitted runs
Relevance labels for validate/test query-document pairs
Evaluation results

Step 1： TOs generate simulated click logs for all training queries

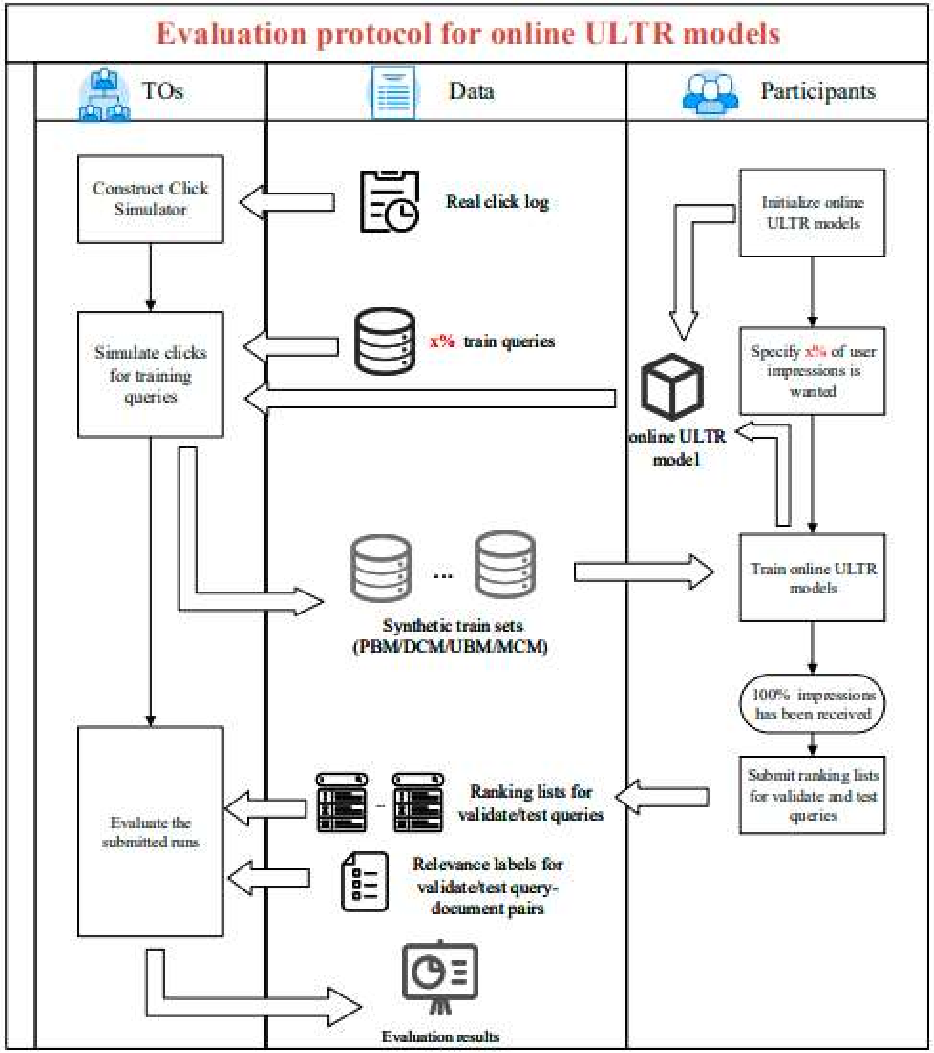
- Use fourt click models (PBM/DCM/UBM/MCM)
- Train and calibrate the click models with real click logs

Step 2： Participants train ULTR 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

Evaluation protocol for online UTLR models


Evaluation protocol for online ULTR models
TOs
Data
Participants
Construct Click Simulator
Real click log
Initialize online ULTR models
Simulate clicks for training queries
x% train queries
Specify x% of user impressions is wanted
online ULTR model
Synthetic train sets (PBM/DCM/UBM/MCM)
Train online ULTR models
100% impressions has been received
Submit ranking lists for validate and test queries
Ranking lists for validate/test queries
Evaluate the submitted runs
Relevance labels for validate/test query-document pairs
Evaluation results

Step 1: Participants submit the ranking lists for training queries

- 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

Experiments

- RQ: Can we evaluate existing ULTR models properly with the ULTRE framework?

Experiment--Evaluating ULTR models with the ULTRE framework

- Dataset (corresponds to the *traditional LTR dataset* in ULTRE framework)
  - Based on SogouSRR [Zhang et al. 2008]
  - 1,211 unique queries with 10 successfully crawled results
  - 1011 for training, 100 for validation and 100 for testing
- Simulation Setup
  - Train LambdaMART with 1% data randomly sampled from the original training set (with 5-level relevance annotations) and use it as the offline production ranker
  - Follow the click-simulation process in the ULTRE framework.
- Model Setup and Evaluation
  - Candidate models: IPW [Joachims et al. 2017] (named PBM-IPS in [Vardasbi et al. 2020]), CM-IPS [Vardasbi et al. 2020] and DLA [Ai et al. 2018].
  - Evaluation metrics: nDCG@5

Evaluation results

	PBM	DCM	UBM	MCM
Production ranker (Baseline)	0.7815			
Full-info (Skyline)	0.8182			
PBM-IPS (IPW)	0.8064	0. 7826	0.8017	0.7647
CM-IPS	0.7894	0.8050	0. 7932	0.7778
DLA	0.8119	0.8173	0.8107	0.7932

Evaluation results

	PBM	DCM	UBM	MCM
Production ranker (Baseline)	0.7815			
Full-info (Skyline)	0.8182			
PBM-IPS (IPW)	**0.8064**	0.7826	0.8017	0.7647
CM-IPS	0.7894	**0.8050**	0.7932	0.7778
DLA	0.8119	0.8173	0.8107	0.7932

When the used behavior models used in click simulation and the correction method of bias are consistent, the results are better than the case in which they don’t agree.

Evaluation results

	PBM	DCM	UBM	MCM
Production ranker (Baseline)	0.7815			
Full-info (Skyline)	0.8182			
PBM-IPS (IPW)	0.8064	0.7826	0.8017	0.7647
CM-IPS	0.7894	0.8050	0.7932	0.7778
DLA	**0.8119**	**0.8173**	**0.8107**	**0.7932**

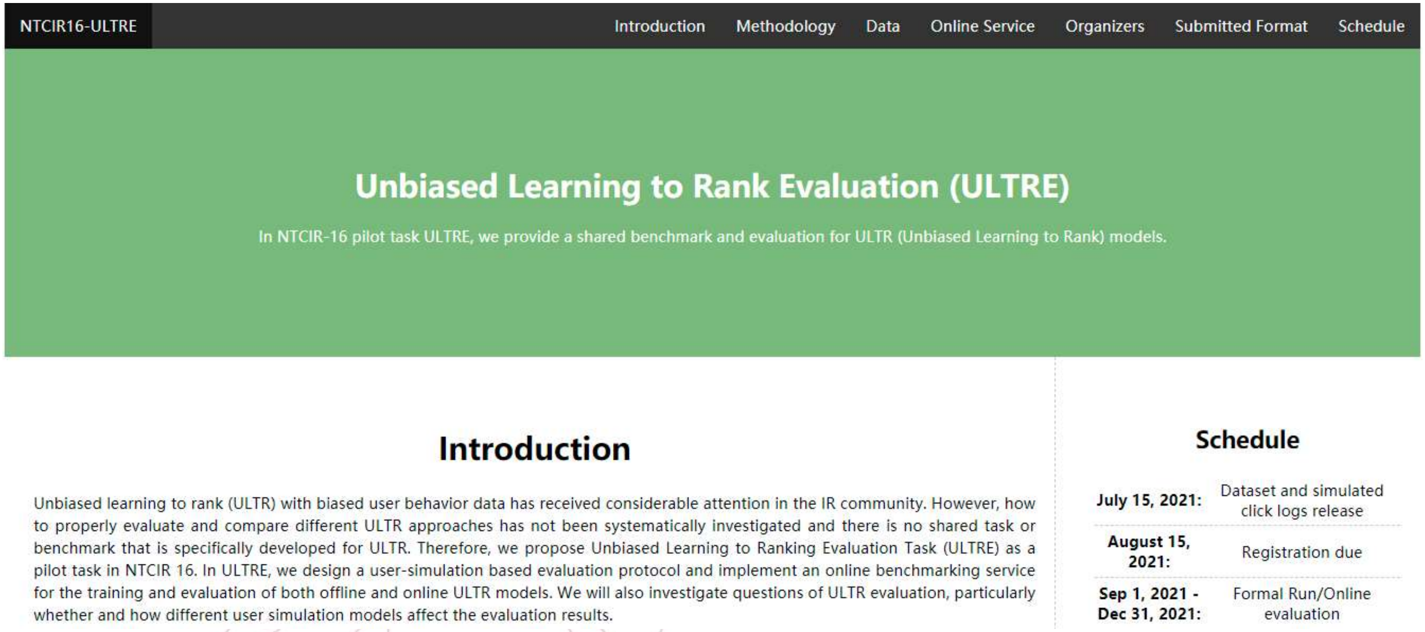
DLA is more robust and more adaptive to the change of user behavior assumption used in the click simulation due to its unification of learning propensity weights (used to correct bias in click data) and leaning ranking models

Conclusion and Future work

- Contribution
  - Propose the ULTRE framework that aims to improve the simulation approach used in previous ULTR evaluation
  - Experiments show that ULTRE framework can provide simulated-based training sets of both quality and diversity and enables us to conduct a thorough and relatively objective comparison of different ULTR models.
- Future work
  - Adopt neural user behavior models such as Context-aware Click Simulator (CCS) [Zhang et al. 2019] for the click simulation
  - Implement online service and Compare online ULTR models under the ULTRE framework

Unbiased Learning to Rank Evaluation Task (ULTRE) in NTCIR-16 (NII Test Collection for IR Systems)

- ULTRE task is a pilot task in NTCIR-16 (http://research.nii.ac.jp/ntcir/ntcir-16/)
- Provide a shared benchmark and evaluation service for ULTR
- NTCIR16-ULTRE task official website: http://ultre.online/



Unbiased Learning to Rank Evaluation (ULTRE)

In NTCIR-16 pilot task ULTRE, we provide a shared benchmark and evaluation for ULTR (Unbiased Learning to Rank) models.

Introduction

Unbiased learning to rank (ULTR) with biased user behavior data has received considerable attention in the IR community. However, how to properly evaluate and compare different ULTR approaches has not been systematically investigated and there is no shared task or benchmark that is specifically developed for ULTR. Therefore, we propose Unbiased Learning to Ranking Evaluation Task (ULTRE) as a pilot task in NTCIR 16. In ULTRE, we design a user-simulation based evaluation protocol and implement an online benchmarking service for the training and evaluation of both offline and online ULTR models. We will also investigate questions of ULTR evaluation, particularly whether and how different user simulation models affect the evaluation results.

Schedule

July 15, 2021:	Dataset and simulated click logs release
August 15, 2021:	Registration due
Sep 1, 2021 - Dec 31, 2021:	Formal Run/Online evaluation

Looking for your participation !

Thanks!

Q&A

Construct Click Simulators

- Use four user simulation models
  - Position-Based Model (PBM) [Craswell et al . 2008]: 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) [Guo et al. 2009]: 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) [Dupret et al. 2008]: a click model that assumes the examination probability on a search result depends on its ranking position and the distance to the lastclicked result.
  - Mobile Click Model (MCM) [Mao et al. 2018]: 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.
- Train and calibrate the user simulation models with real query logs

Experiments

- RQ : How do different click simulators performs in predicting clicks and generating synthetic click logs?

Examining click simulators

- Dataset (corresponds to the *Real click logs* in the ULTRE framework)
  - real search log dataset released by Chinese commercial search engine Sogou.com (1.6 million sessions, 1211 unique queries)
- Evaluation metrics
  - LogLikelihood (LL) and Perplexity (PPL)
  - Reverse/Forward PPL [Dai et al. 2021]

Performance on predicting clicks

	LL	PPL
DCM	–0.1848	1.2363
PBM	–0.1721	1.2059
UBM	–0.1513	1.2029
MCM	**–0.1503**	**1.1787**

Quality of generated click logs

Reverse PPL: the PPL of a surrogate model (an intermediary to evaluate the similarity between the generated samples and the real data samples) that is trained on generated samples and evaluate on real data.

Forward PPL: the PPL of a surrogate model that is trained on real data and evaluated on generated samples

	Surrogate DCM		Surrogate PBM		Surrogate UBM		Surrogate MCM	
Real data	1.2363	1.2363	1.2059	1.2059	1.2029	1.2029	1.1787	1.1787
DCM samples	–	–	1.2374	1.3688	1.2350	1.3625	1.2191	1.3137
PBM samples	1.2824	1.2272	–	–	1.2061	1.1880	1.2053	1.2152
UBM samples	1.2409	1.2317	1.2055	1.1953	–	–	1.1802	1.1764
MCM samples	1.2388	1.2248	1.2061	1.1841	1.2031	1.1831	–	–