Baidu-ULTR: a large-scale dataset for Unbiased Learning to Rank

Haitao Mao

Joint work with Lixin Zou, Xiaokai Chu, Jiliang Tang, Shuaiqiang Wang, Wenwen Ye, Changying Hao, Dawei Yin.

Outline

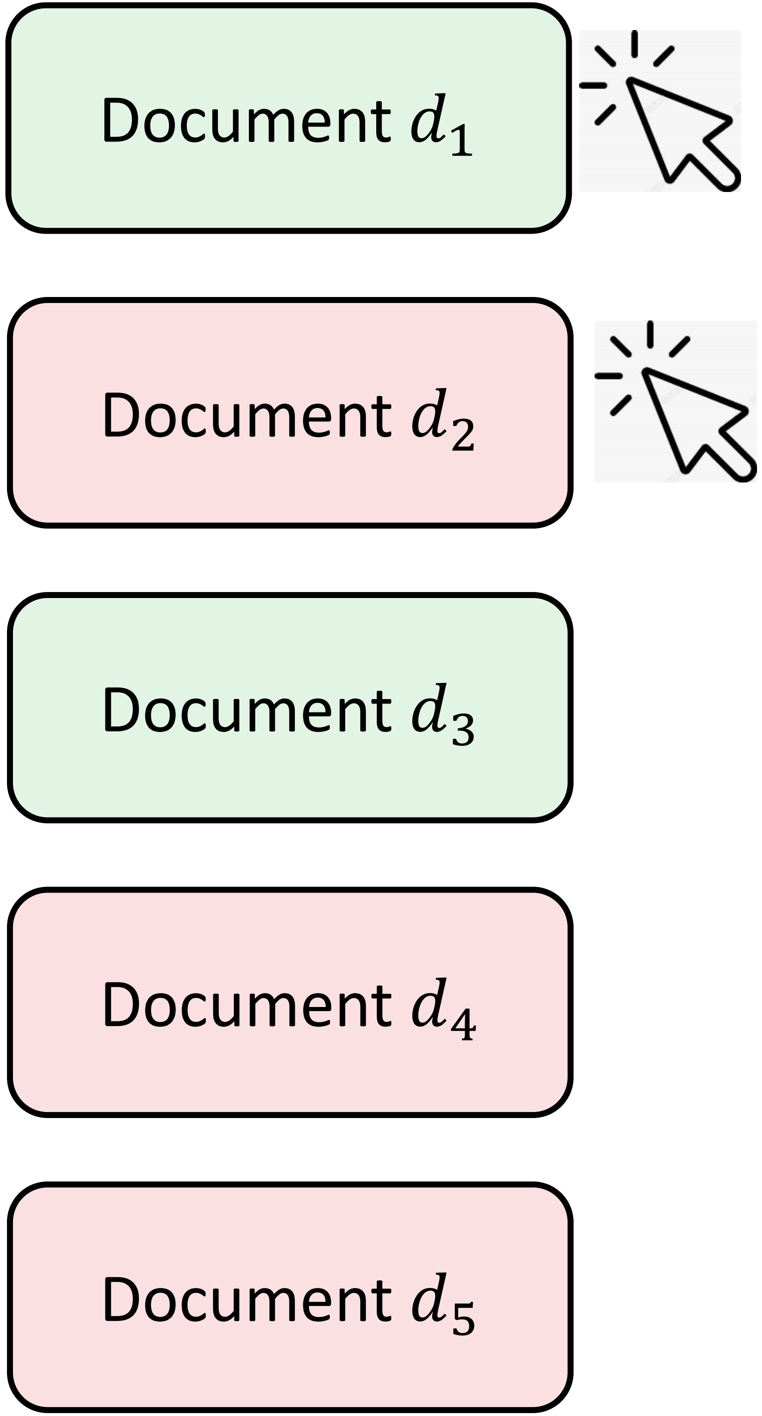
- A brief introduction on Unbiased Learning to Rank (ULTR)
- Real-world challenge & our dataset
- WSDM Cup tasks introduction
- Experiment & data analysis
- Discussion

A brief introduction on ULTR

Brief Introduction

- **Learning to Rank**
  rank the document with higher relevance to query higher position

- **Unbiased Learning to Rank**
  Learn an ideal relevance model with biased click model



Real-world Challenge & our dataset

What we want toward an ideal dataset

- ☐ The dataset more like the real-world scenario
- ☐ The training and evaluation procedure similar with the real-world scenario
- ☐ The dataset can allow us utilize the advanced techniques

Dataset more like real-world scenario

- Previous datasets only provides position, the only one page presentation feature.
- The new modern search engine can provide more page presentation features

(a) Rich Page Presentation Information in Baidu-ULTR

Practical train and evaluation prototype

Table 1: Characteristics of publicly available datasets for unbiased learning to rank

	Training Implicit Feedback Data					Validation & Test Data				
Dataset	# Query	# Doc	# User Feedback	# Display-info	# Session	# Query	# Doc	# Label	# Feature	Pub-Year
Yahoo Set1	19,944	473,134	1 (Simulated click)	1 (Position)	-	9,976	236,743	5	519	2010
Yahoo Set2	1,266	34,815	1 (Simulated click)	1 (Position)	-	5,064	138,005	5	596	2010
Microsoft	≈18,900	≈2,261,000	1 (Simulated click)	1 (Position)	-	≈12,600	≈1,509,000	5	136	2010
Istella	23,219	7,325,625	1 (Simulated click)	1 (Position)	-	1,559	550,337	5	220	2016
Tiangong	3,449	333,813	1 (Real Click)	1 (Position)	3,268,177	100	10,000	5	33	2018
Baidu	383,429,526	1,287,710,306	18 (Real Feedback)	8 (Display Info)	1,210,257,130	7,008	367,262	5	ori-text	2022

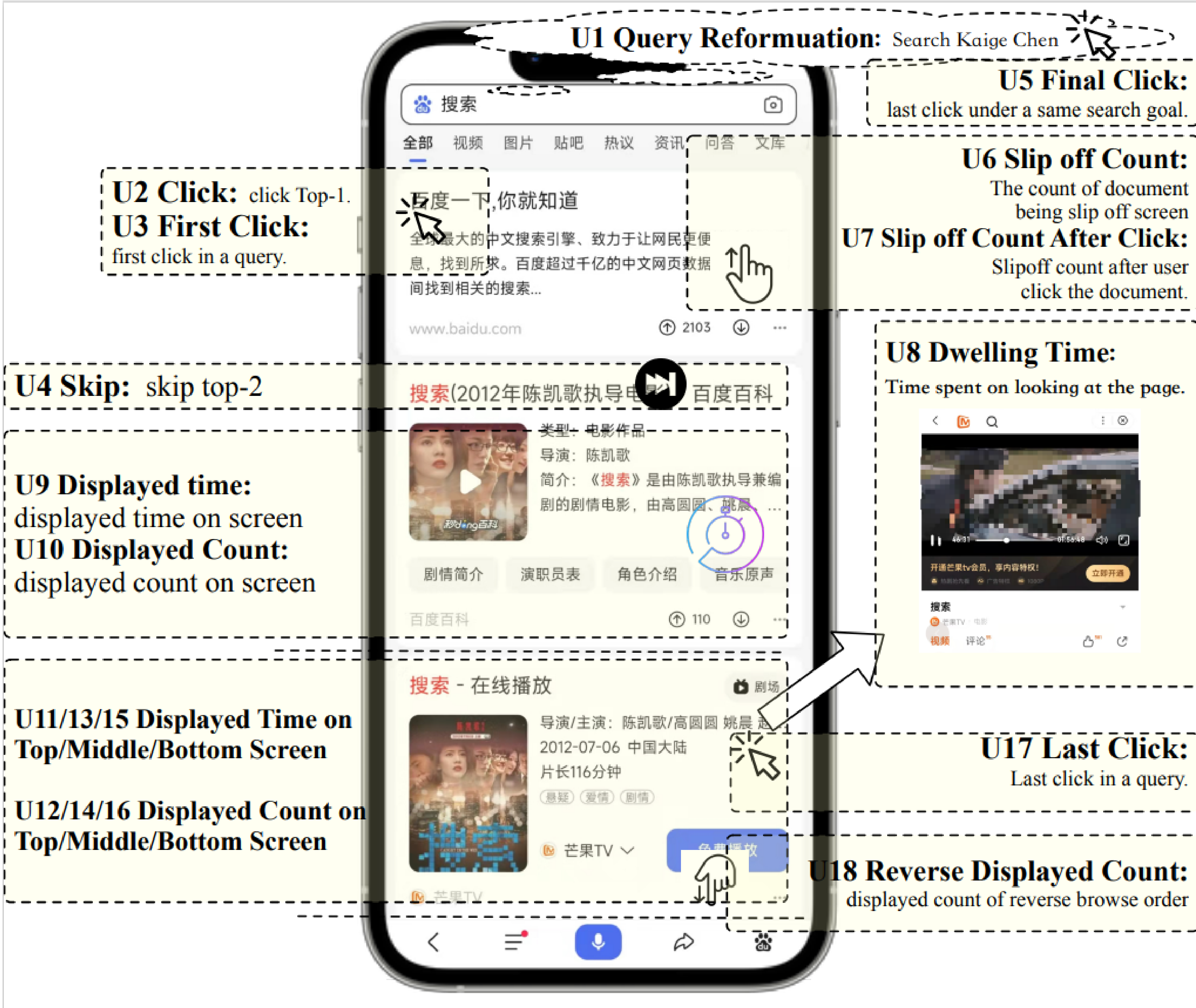
- Pipeline: (1) click data for training (2) annotation data for evaluation
- Existing datasets utilize synthetic data for training, and small annotation set
- Provide real-world click data and a fairly large testset

Utilize more advanced techniques

- Large-scale pretrain model, e.g., BERT, ERNIE, are common utilized in Natural Language Processing.
- Existing datasets provide only provide preprocess features, e.g., tf-idf, BM25
- Baidu-ULTR provides raw tokens after desensitization.
- The dataset size is 20 times larger than existing datasets.

Go beyond simple ULTR

More user behavior:
Click may not be the only signal for ULTR



(b) Rich User Behaviors in Baidu-ULTR

Tasks

Task introduction

- Unbiased learning to rank

  Click data for training and Expert Annotation dataset for evaluation

- Pre-training for Web search

  Click and part of the part of the Expert Annotation dataset for training.
  Expectively, click for pretrain, Expert Annotation for finetuning

No Expert Annotation data for training in ULTR task!

Task submission guidance

- Download the test dataset on the official website.
- Put the test data into path “test_annotate_path”, then run our submit script and load the model you saved.
- Submit your predict result on the platform.

Experiments & Data Analysis

Primary Experiments

Table 4: Comparison of unbiased learning to rank (ULTR) algorithms with different learning paradigms on Baidu-ULTR using cross-encoder as ranking models. The best performance is highlighted in bold

	DCG@1	ERR@1	DCG@3	ERR@3	DCG@5	ERR@5	DCG@10	ERR@10
Naive	1.235±0.029	0.077±0.002	2.743±0.072	0.133±0.003	3.889±0.087	0.156±0.003	6.170±0.124	0.178±0.003
IPW	1.239±0.038	0.077±0.002	2.742±0.076	0.133±0.003	3.896±0.100	0.156±0.004	6.194±0.115	0.178±0.003
REM	1.230±0.042	0.077±0.003	2.740±0.079	0.132±0.003	3.891±0.099	0.156±0.004	6.177±0.126	0.178±0.004
PairD	1.243±0.037	0.078±0.002	2.760±0.078	0.133±0.003	3.910±0.092	0.156±0.003	6.214±0.114	0.179±0.003
DLA	**1.293**±0.015	**0.081**±0.001	**2.839**±0.011	**0.137**±0.001	**3.976**±0.007	**0.160**±0.001	**6.236**±0.017	**0.181**±0.001

- No algorithm shows much better result than the naïve algorithm
- DLA perform best across all methods

Performance on query with different frequency

Table 5: Performance comparison of evaluation ULTR algorithms versus different search frequencies. The best performance is highlighted in boldface.

Model	DCG@3			DCG@5			DCG@10		
	High	Mid	Tail	High	Mid	Tail	High	Mid	Tail
Naive	3.960±0.058	2.992±0.119	1.742±0.079	5.596±0.098	4.254±0.142	**2.474**±0.092	8.812±0.140	**6.777**±0.173	3.942±0.121
IPW	4.017±0.132	2.976±0.111	1.722±0.061	5.699±0.145	4.235±0.140	2.447±0.090	8.969±0.146	6.762±0.163	3.925±0.109
REM	3.994±0.114	2.982±0.124	1.723±0.067	5.665±0.128	4.237±0.158	2.454±0.074	8.904±0.147	6.755±0.183	3.927±0.104
PairD	4.018±0.102	2.993±0.110	**1.750**±0.079	5.662±0.120	4.267±0.129	2.474±0.088	8.924±0.145	6.804±0.153	**3.961**±0.119
DLA	**4.226**±0.042	**3.073**±0.022	**1.750**±0.016	**5.894**±0.030	**4.300**±0.020	2.472±0.009	**9.147**±0.044	6.767±0.027	3.920±0.009

- All algorithms performance drop from high to tail
- Naïve algorithm shows good performance in Tail query

Data Analysis – Expert Annotation

Table 3: Distribution of Relevance Label.

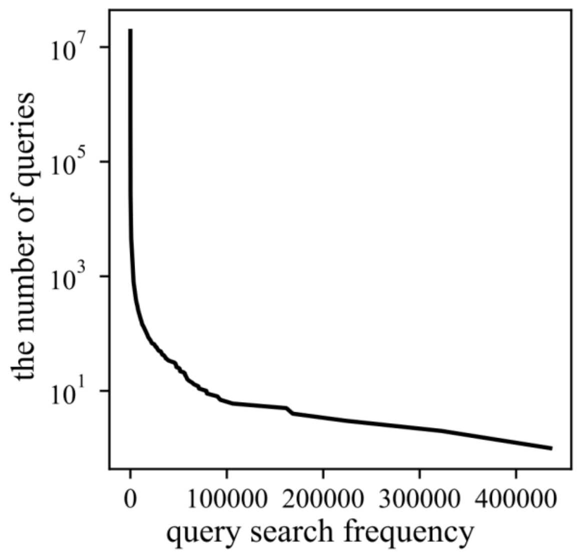
Grade	Label	# Query-Doc	Ratio of Label
Perfect	4	714	1.80%
Excellent	3	28,172	9.21%
Good	2	112,759	28.36%
Fair	1	36,622	9.21%
Bad	0	219,305	55.16%

- Perfect only occupies 1.8%
- Bad documents take over 50% document

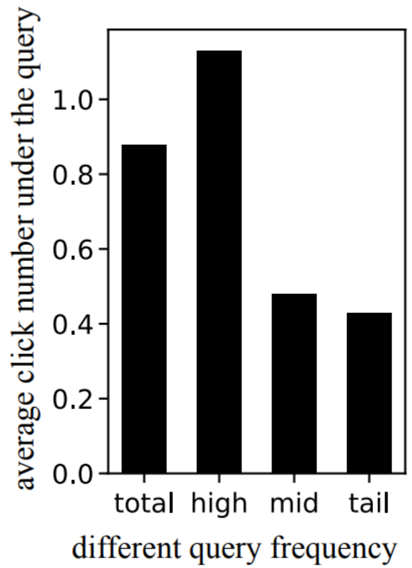
Table 4: The general guideline of annotation.

Label	Guideline
0 (bad)	Useless or outdated documents that do not meet the requirements at all.
1 (fair)	Helpful to some extent but deficient in authority, timeliness document.
2 (good)	Meet the requirement of the query.
3 (excellent)	Meet the requirement of the query and timeliness document.
4 (Perfect)	Meet the requirement of the query, timeliness, and authoritative document.

Data Analysis – Long-tail distribution



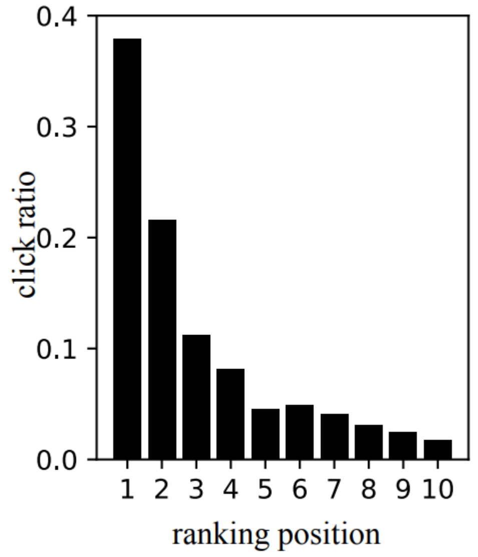
(a) the distribution of query frequency



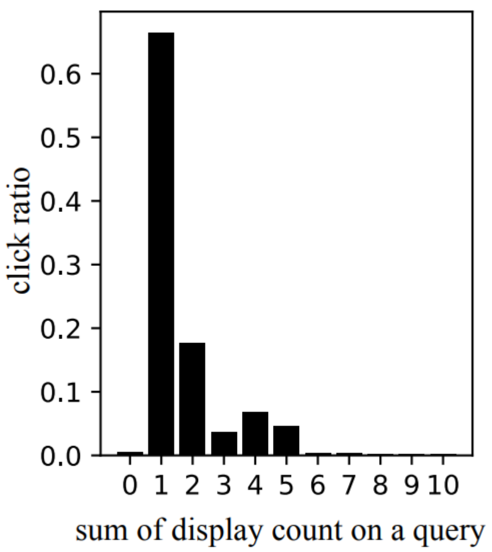
(b) average click number of query under different frequency

Long tail distribution appears in many behavior. For example, Over 60% searches are based on top 10% high frequency.

Data Analysis – click, query, displayed features



(c) click ratio on different ranking position



(d) click ratio vs query displayed count

- Click shows strong correlation with displayed feature.
- Click shows correlation with query frequency.

Discussion

-- Challenge & Opportunity

Challenge

- Biases in Real-World User feedback

- Long-tail Phenomenon

- Mismatch between Training and Test
  - In training stage, only top-10 documents recorded.
  - In test stage, top-30 documents and further documents samples

Opportunity

- Pretraining models for Ranking
- Causal Discovery
- Multi-task Learning

Thanks!