2023 WSDM CUP 排序学习赛题详解

10月19日
20:00-21:00

AI TIME
Artificial Intelligence Time

关于我们

AI TIME源起于2019年，旨在发扬科学思辨精神，邀请各界人士对人工智能理论、算法和场景应用的本质问题进行探索，加强思想碰撞，链接全球AI学者、行业专家和爱好者，希望以辩论的形式，探讨人工智能和人类未来之间的矛盾，探索人工智能领域的未来。

三年来，AI TIME已经邀请了800多位海内外讲者，举办了逾400场活动，超400万人次观看，汇集了全球逾百位志愿者团队；活动内容涵盖了AI领域的所有方向和AI+交叉领域的思想碰撞，主办与参与了丰富的多种活动形式：如世界人工智能大会、智源大会、政府产业论坛、大佬思辨、博士思辨、走进高校、技术分享、闭门脑暴会等。

AI TIME丰富的内容专题

专题	简介
AI TIME Debate	邀请AI领域的顶级专家，针对前沿问题，对AI理论、算法与应用场景等进行深度探索与思辨。
AI TIME PhD Debate	邀请海内外AI方向PhD，针对AI前沿学术问题，进行交流、探索与思想的碰撞。
AI TIME PhD	邀请在AI顶会发表论文的PhD分享最前沿的研究成果，对相关学术议题进行交流探讨。
AI TIME 青年科学家	邀请海内外AI领域青年学者分享最新科研创新成果，促进青年人才的思维碰撞，搭建一个高质量的青年AI人才交流平台。
AI TIME 走进高校	邀请AI领域大佬一起走进国内各高校，多角度地探索AI的研究与应用，为高校和产业界建立一个共享、开放的平台。
AI TIME Master	邀请世界级大师来做线上talk，与大家分享世界级科学家的前瞻性观点。
AI TIME 科普小课堂	针对大众关注的内容做科普分解，让更多的人了解AI,了解前沿科技。
百辨太魔人	邀请AI TIME的志愿者通过思辨的形式，以太魔人的真实经历为大家传递正能量。干货、有趣、接地气、共鸣，邀请观众与太魔人一起成长
科普大佬说	邀请AI 学术界与产业界的大佬针对人工智能领域的热点问题进行知识科普，追溯科学本质，探索科技前沿，带领大家一起走进人工智能的世界。

AI TIME
Artificial Intelligence Time

- 想与讲者和相关领域的优秀研究人员近距离交流吗？可以！请添加我们的“**小A助手**”（微信号AITIME_HY），回复关键词“PhD”即可拉您入群。
- 请大家关注AI TIME 官方公众号“**AI TIME论道**”（微信号lundaoAI）及微博账号@AITIME论道，更有超多报告PPT、往期精彩AI TIME Debate，AI TIME PhD学术分享的视频和独家资源报道等待小伙伴去挖掘宝藏哟~

添加“小A助手”
回复关键词“PhD”入群

关注公众号
“AI TIME论道”

关注微博账号
@AITIME论道

AI TIME志愿者——我们都是太魔人

- AI TIME的志愿者团队来自五湖四海。他们就读或毕业于海内外高校：清华大学、北京大学、北航、帝国理工、密歇根大学…一起带来多元的文化视角。在不同之外，Ta们共享着同一个理念：尊重科学、尊重思辨、尊重每一个人。

- AI TIME志愿者团队欢迎关注AI的有活力和精力的小伙伴加入我们来共享共建这个AI乌托邦。在这里，你将有机会与世界级的AI专家当面请教，与一流学府的精英学子互相学习，与一线企业或独角兽中的技术大拿分享心得。

欢迎加入AI TIME，遇见太魔人

AI TIME志愿者招募

Volunteer Recruitment

优秀团队

AI TIME志愿者团队分别来自清华、北大、北邮、北外、北林等顶尖高校。
我们欢迎每一位AI爱好者的加入！

AI TIME
部分志愿者合影

感谢AI TIME大家庭中的每一位讲者与志愿者！

欢迎每一位喜爱AI的小伙伴加入我们呀！

本期讲者

邹立新，百度大搜资深算法工程师，2020年博士毕业于清华大学计算机系，同年AIDU计划加入百度，在上层排序组先后负责大搜深度语言匹配和点击率预估等核心方向。目前，在相关方向顶级学术会议上发表学术论文20余篇，获2019年KDD Cup Runner Up Award。

毛海涛，密歇根州立大学一年级博士生，师从汤继良教授，曾以一作身份获CIKM2021 best short paper。

Unbiased Learning & Pre-training for Web Search

Lixin Zou, Haitao Mao

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

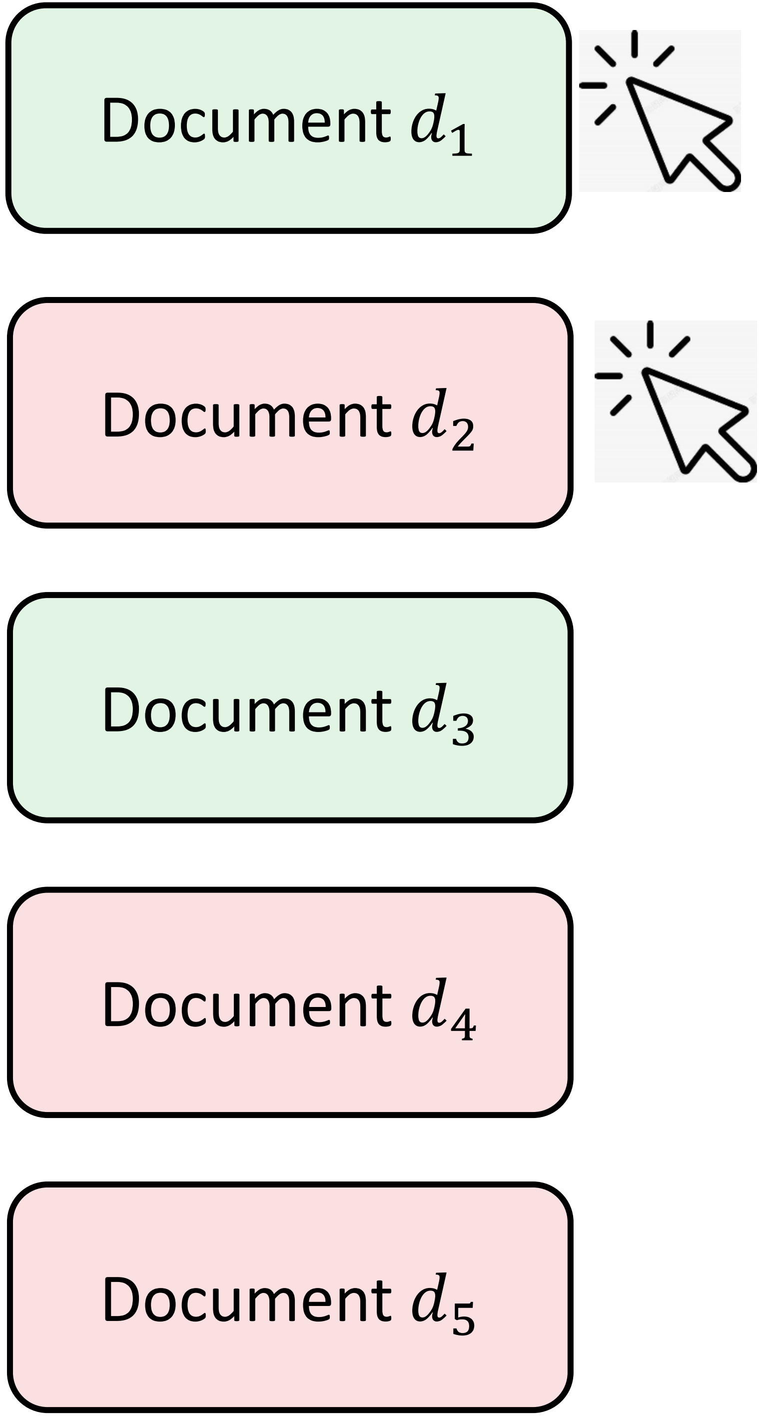
Outline

- A brief introduction on Learning to Rank tasks
- Dataset introduction
- Task submission guidance
- Experiment & data analysis & Further discussion

A brief introduction on learning to rank tasks

Brief Introduction

- **Learning to Rank (LTR)**
  rank the document with higher relevance to query higher position
- **Unbiased Learning to Rank (ULTR)**
  Learn an ideal relevance model with biased click model
- **Pre-train Language Model**
  Learn a relevance model with the help of pretraining.



Challenge in ULTR

- The user behavior is complicated rather than just simple click.
- The user behavior can be affected by different display feature rather than just simple position.
- Existing academic datasets only consider position and click.

Challenge in pre-training

- No dataset provides the raw text information, only processed feature

- Directly apply recent advancements in PLMs to web-scale search engine systems since explicitly capturing the comprehensive relevance between queries and documents is crucial to the ranking task.

- How can a ranking-model query-document relevance relationship is an non-trivial problem

Dataset Introduction

What we want toward an ideal dataset

- The training and evaluation procedure similar with the real-world scenario

- The dataset more like the real-world scenario

- The dataset can allow us utilize the advanced techniques

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

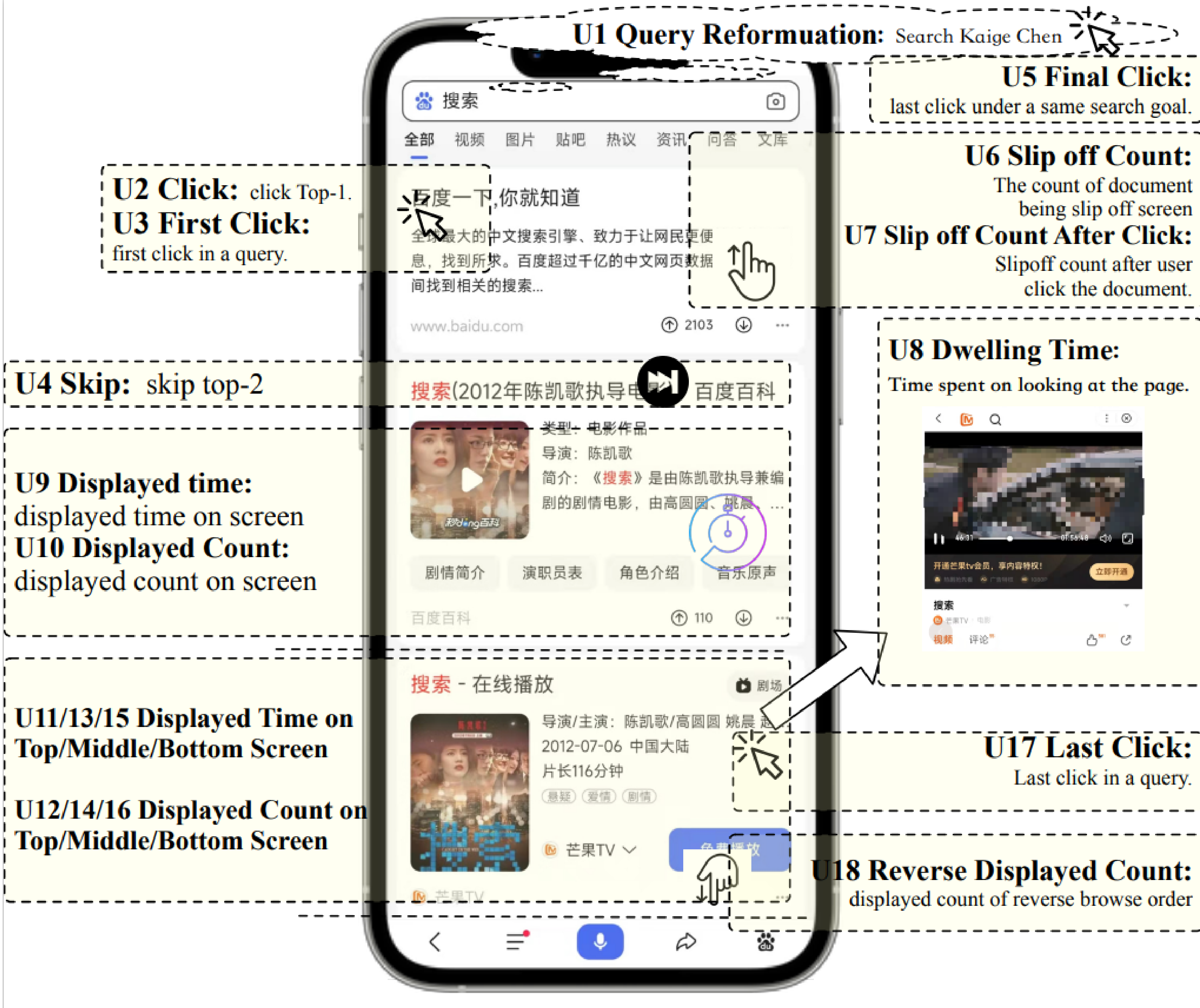
Dataset more like real-world scenario

- 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

Dataset more like Real-world scenario

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



(b) Rich User Behaviors in Baidu-ULTR

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.

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

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 & Further Discussion

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.

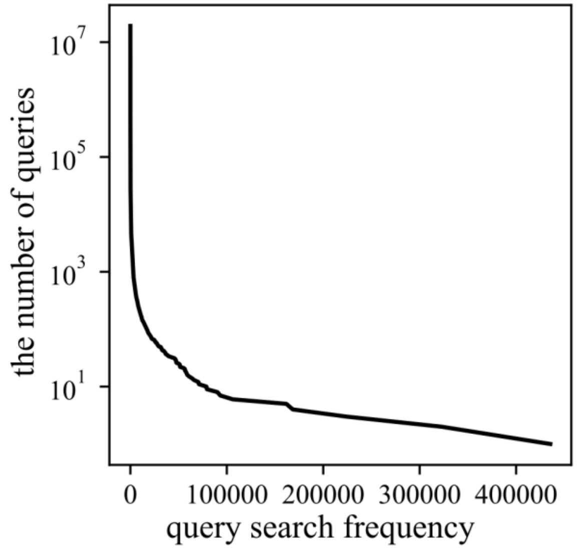
Grade	Label	# Query-Doc	Ratio of Label
Perfect	4	714	1.80%
Excellent	3	28,172	9.21%
Good	2	112,759	28.36%
Fair	1	36,622	9.21%
Bad	0	219,305	55.16%

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

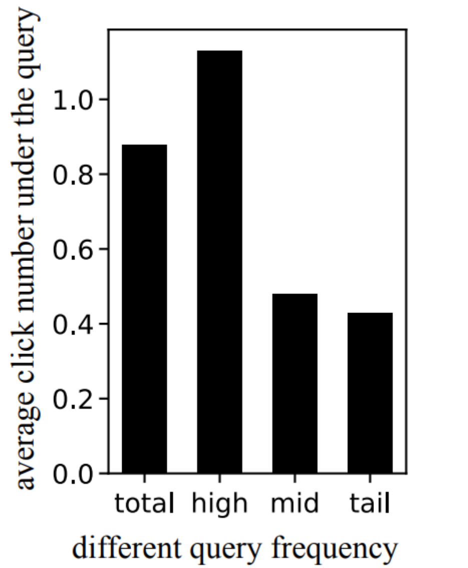
Table 4: The general guideline of annotation.

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



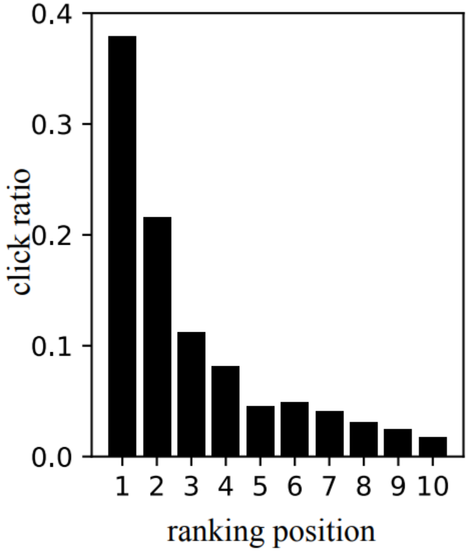
(a) the distribution of query frequency



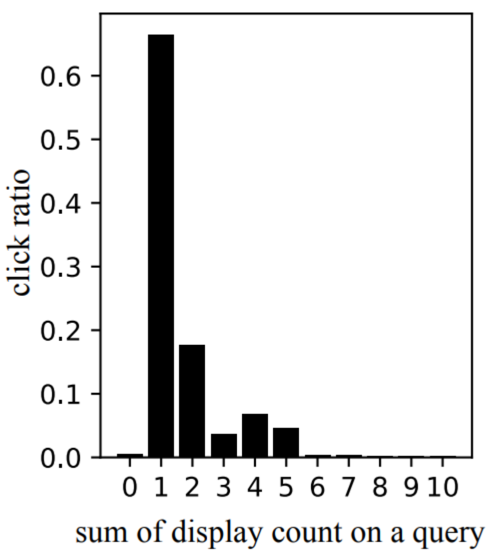
(b) average click number of query under different frequency

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.

Further Discussion

- 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

Thanks!