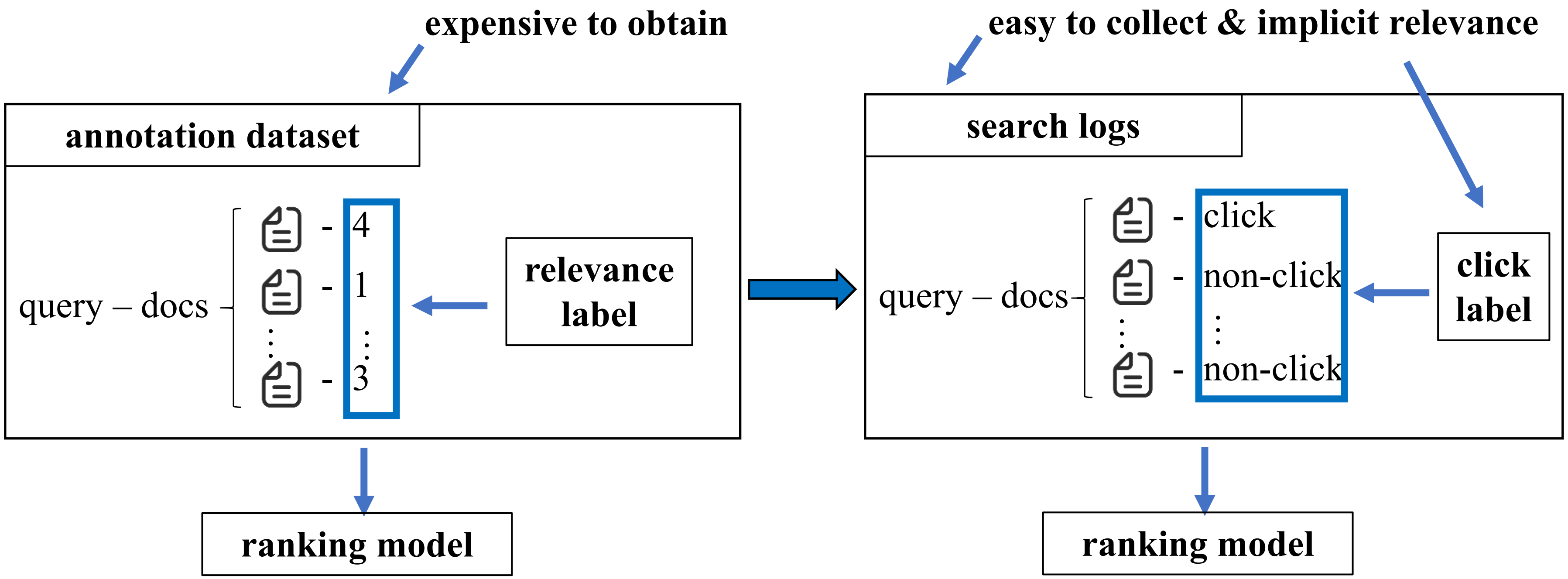
CIR at the NTCIR-17 Unbiased Learning to Rank Evaluation Task 2 (ULTRE-2)

Lulu Yu, Keping Bi, Jiafeng Guo, Xueqi Cheng
Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS)
CAS Key Lab of Web Data Science and Technology

Background

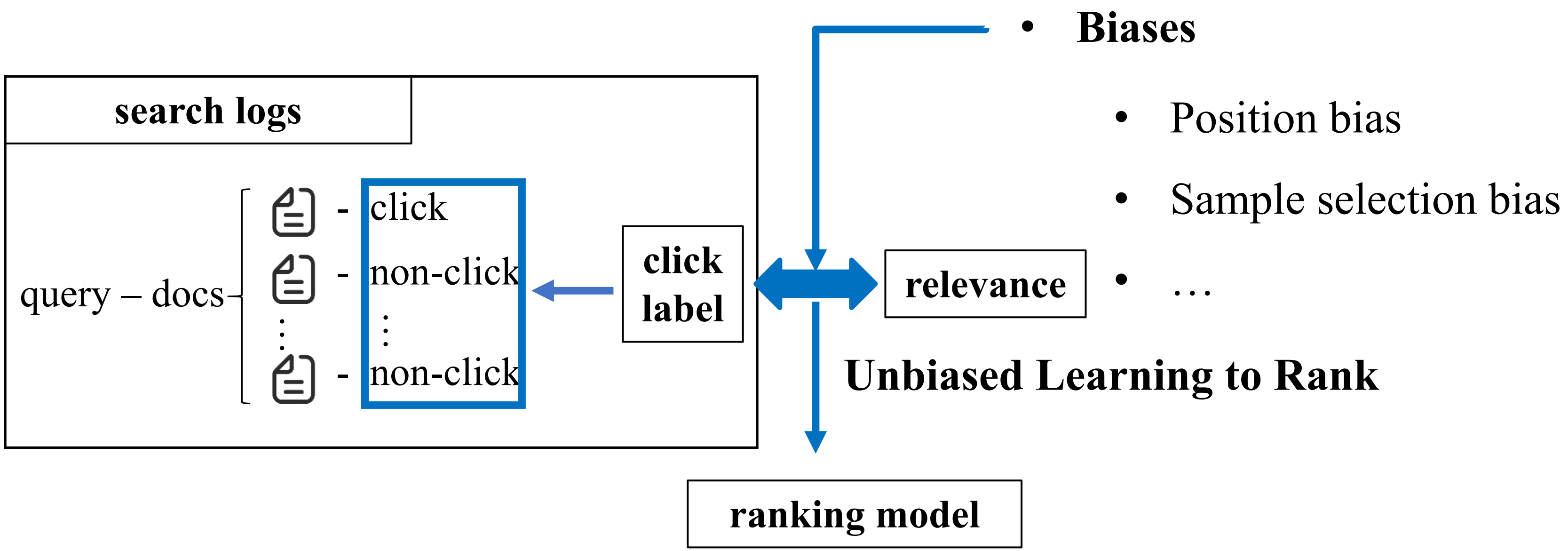
Background

- **Learning to Rank**



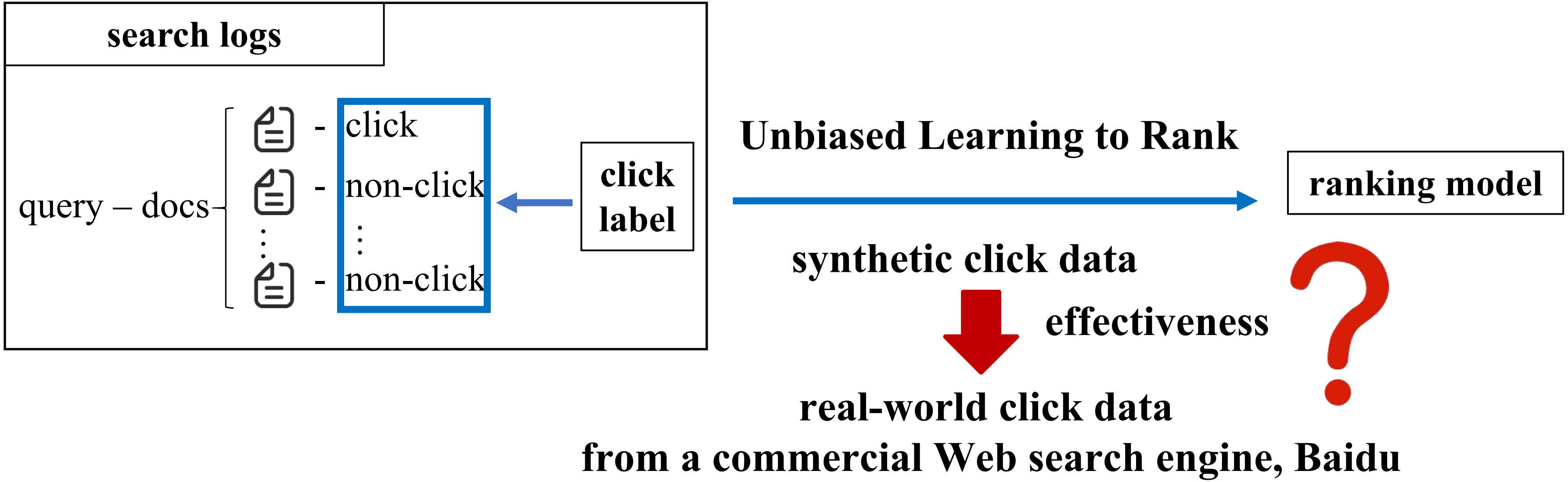
Background

- **Unbiased Learning to Rank**



Background

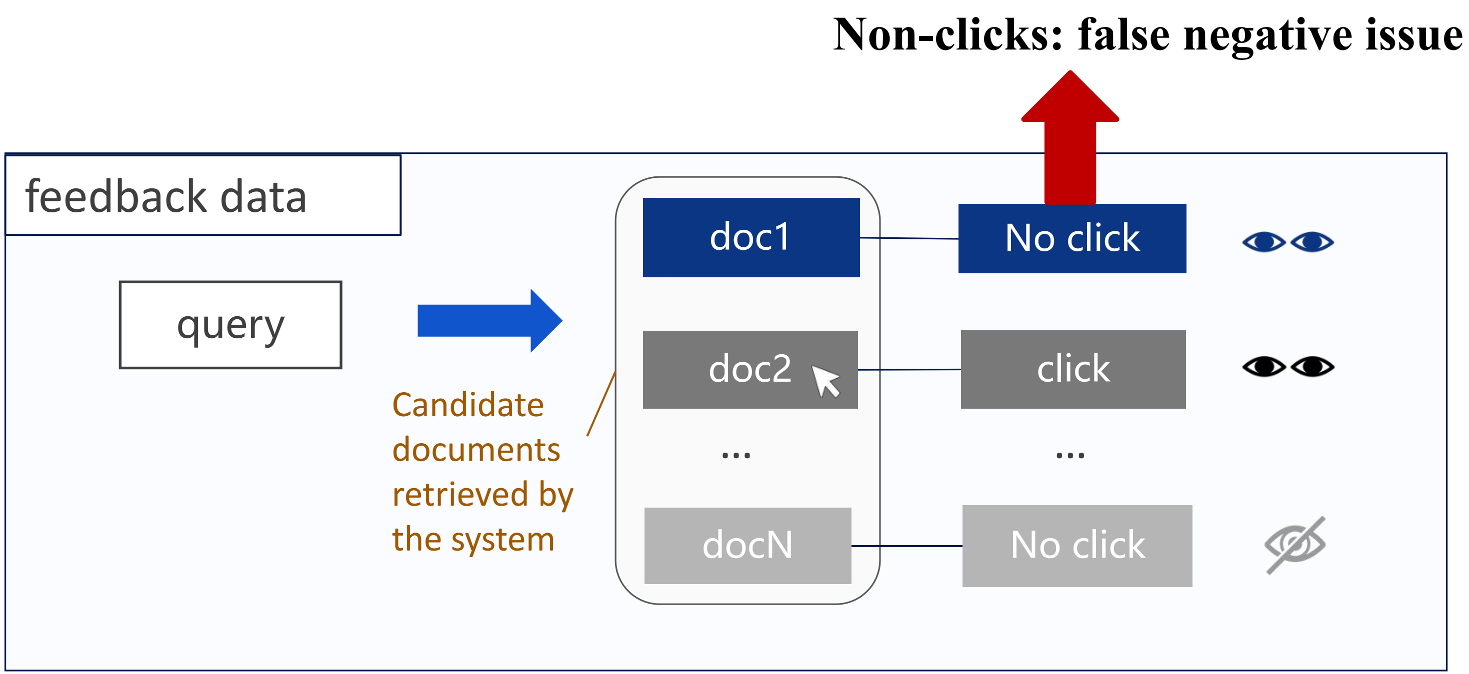
- **NTCIR-17 ULTRE-2 task**



Motivation & Methods

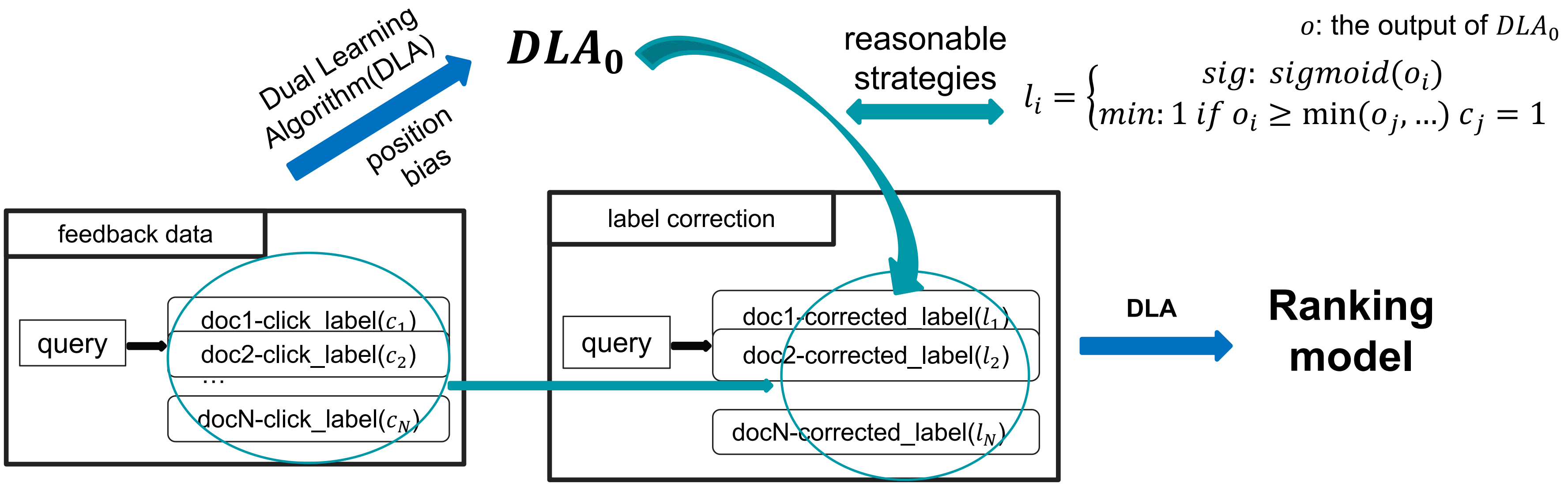
Motivation

- **non-clicks do not mean irrelevant ⬄ false negative issue**



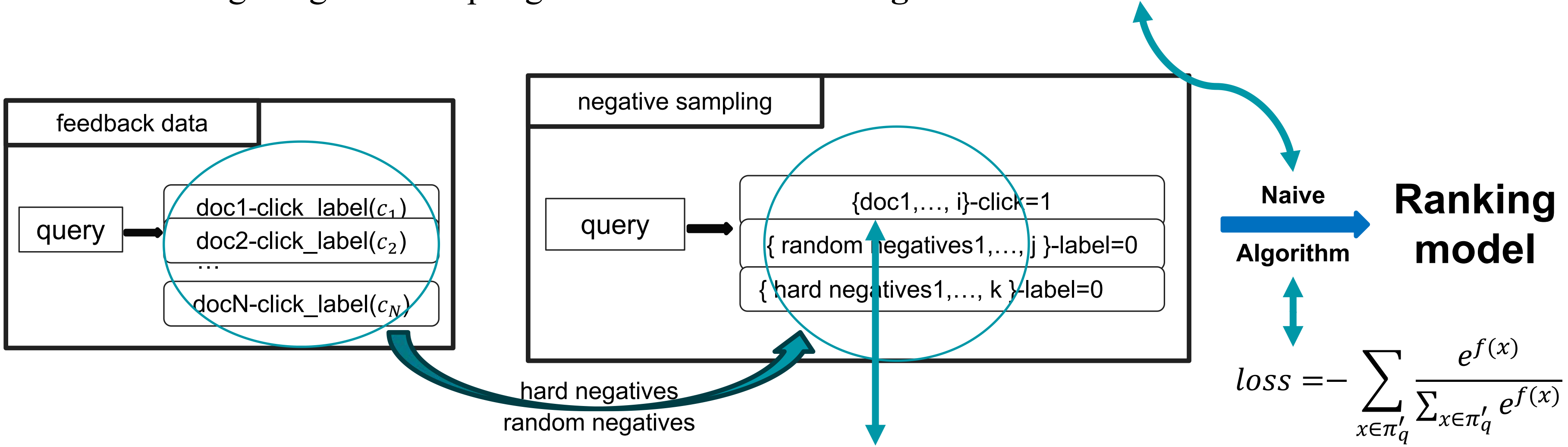
Methods

- **Label Correction**
  - ➢ correct the labels for **non-clicked** items by a relevance judgment model trained from DLA



Methods

- **Negative Sampling**
  - ➢ through negative sampling → **reconstruct the original result lists**



$$loss = -\sum_{x \in \pi'_q} \frac{e^{f(x)}}{\sum_{x \in \pi'_q} e^{f(x)}}$$

  - ➢ "click-only" scheme: preserve clicked results
  - ➢ "last-click" scheme: preserve all the results before the last clicked result

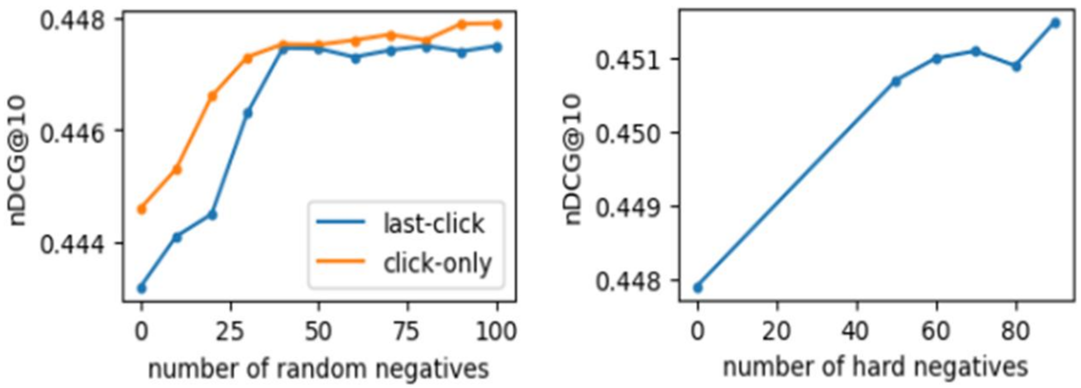
Experiments & Results

Experiments

- **Experimental implementation**
  - **Input features**
    - for label correction method
      - traditional features of 13 dimensions
      - pretrained score
    - for negative sampling method
      - extracted traditional word matching features (e.g. LM-DIR, BM25) of 24 dimensions
  - **Model architecture**
    - Feature projection: project features to a higher dimension
    - Ranking model: a deep neural network with three hidden layers

Results

- **Effect of Negative Sampling**
  - We investigate the use of negative sampling on the validation set.
  - The nDCG@10 on the validation set indicates that this approach is effective in improving performance.



Performance curves of two schemes ("click-only" and "last-click") w.r.t. the number of random and hard negatives. (a) Performance curves of two schemes w.r.t. the number of random negatives. (b) The Performance curve of the "click-only" scheme w.r.t. the number of hard negatives.

Results

- **Effect of Label Correction**
  - DLA with Label Correction **outperforms** the basic DLA model, under various strategies.
  - The underline denotes the performance of the baseline DLA.

Model	nDCG@10			DCG@10
	test	total valid	20% valid	test
Scratch-DLA-LC (sig)	**0.5355**	**0.5019**	/	11.4538
Aux-DLA-LC (sig)	0.5326	0.5015	/	11.3898
Scratch-DLA-LC (min)	unk	0.4816	/	unk
Aux-DLA-LC (min)	unk	0.4947	/	unk
DLA	0.5247	0.4920	/	11.2031
lgbBase	0.5350	/	0.5003	11.4794
lgbBaseAdd	0.5333	/	0.5021	11.4616

Conclusion

Conclusion

- We focus on the false negative issue and propose two approaches to tackle this issue: label correction and negative sampling.
- Both methods can enhance the model performance and our best method (label correction) has achieved nDCG@10 of 0.5355, which is 2.66% better than the best score from the organizer.

Model	nDCG@10			DCG@10
	test	total valid	20% valid	test
Scratch-DLA-LC (sig)	**0.5355**	**0.5019**	/	11.4538
Aux-DLA-LC (sig)	0.5326	0.5015	/	11.3898
Scratch-DLA-LC (min)	unk	0.4816	/	unk
Aux-DLA-LC (min)	unk	0.4947	/	unk
DLA	0.5247	0.4920	/	11.2031
lgbBase	0.5350	/	0.5003	11.4794
lgbBaseAdd	0.5333	/	0.5021	11.4616

Team	Submitted run	nDCG@10
Organizer	click-point	0.3326
	click-pair	0.5100
	click-softmax	0.5144
	IPS-PBM	0.5199
	IPS-DCM	0.5131
	IPS-UBM	0.4875
	DLA-PBM	0.5216
	DLA-DCM	0.5199
	DLA-UBM	0.5196
	PRS	0.4970

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

Thanks!

yululu23s@ict.ac.cn

NTCIR