NTCIR-16

OVERVIEW OF THE NTCIR-16 SESSION SEARCH (SS) TASK

Jia Chen[†], Weihao Wu [†], Jiaxin Mao[‡], Beining Wang[†], Fan Zhang[†], Yiqun Liu*[†]

† Department of Computer Science & Tech., Tsinghua University

‡ Gaoling School of AI, Renmin University of China

chenjia0831@gmail.com, yiqunliu@tsinghua.edu.cn

Introduction

Multi-query sessions/tasks are common in search process nowadays.

- Considering the **contextual information** within sessions has been proved efficient for user intent modeling in IR communities.
- Existing tasks in NTCIR have not involved **session-based retrieval** yet.

Existing relevant tasks

- TREC Session Track 2011-2014: small-scale datasets, simulated search tasks, only evaluate the last query;
- TREC Dynamic Domain (DD) Track 2015-2017: user simulators, ignore user reformulations.

==> NTCIR-16 Session Search Task

Introduction

To better assess the search effectiveness at both query-level and session-level, we aim to set up two subtasks as follows:

***Fully Observed Session Search (FOSS)*:**

- For a k-length session, we provide full session contexts in the first $(k-1)$ queries.
- NDCG, AP, RBP

Partially Observed Session Search (POSS):

- In this subtask, we truncate all sessions before the last query. For a session with k queries $(k \geq 2)$, we only reserve the session contexts in the first n queries, where $1 \leq n \leq k-1$.
- RS-DCG, RS-RBP

Schedule

Table 1: NTCIR-16 SS timeline (time zone: AOE).

Session Dataset Release	Aug 15, 2021
Formal Run	Oct 1 - Dec 31, 2021
Preliminary Evaluation	Dec, 2021
Relevance Assessment	Feb, 2022
Evaluation Results Release	Mar, 2022

Participation

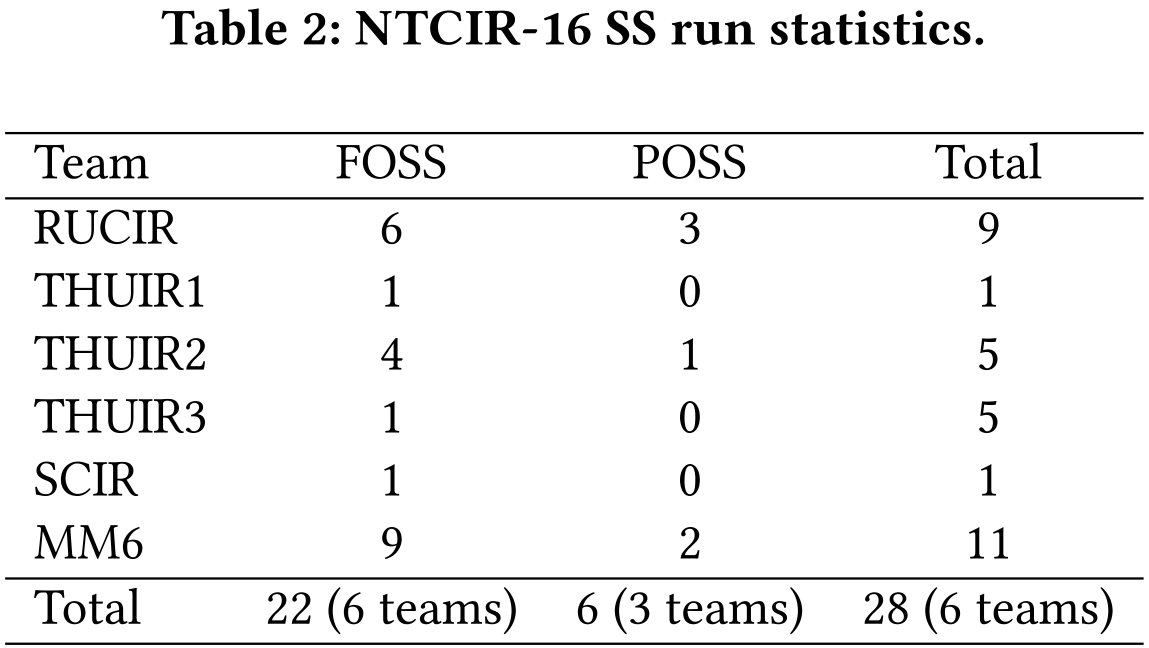


Table 2: NTCIR-16 SS run statistics.

Team	FOSS	POSS	Total
RUCIR	6	3	9
THUIR1	1	0	1
THUIR2	4	1	5
THUIR3	1	0	5
SCIR	1	0	1
MM6	9	2	11
Total	22 (6 teams)	6 (3 teams)	28 (6 teams)

Dataset & Resources

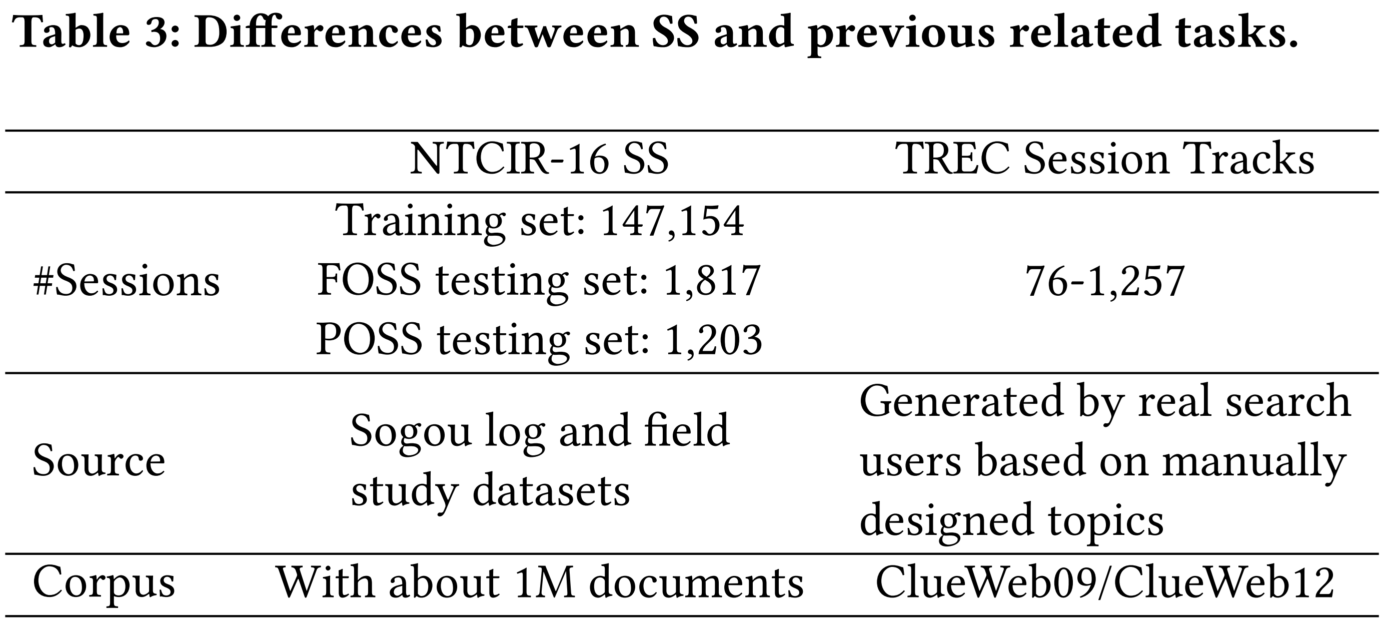


Table 3: Differences between SS and previous related tasks.

	NTCIR-16 SS	TREC Session Tracks
#Sessions	Training set: 147,154 FOSS testing set: 1,817 POSS testing set: 1,203	76-1,257
Source	Sogou log and field study datasets	Generated by real search users based on manually designed topics
Corpus	With about 1M documents	ClueWeb09/ClueWeb12

Dataset & Resources

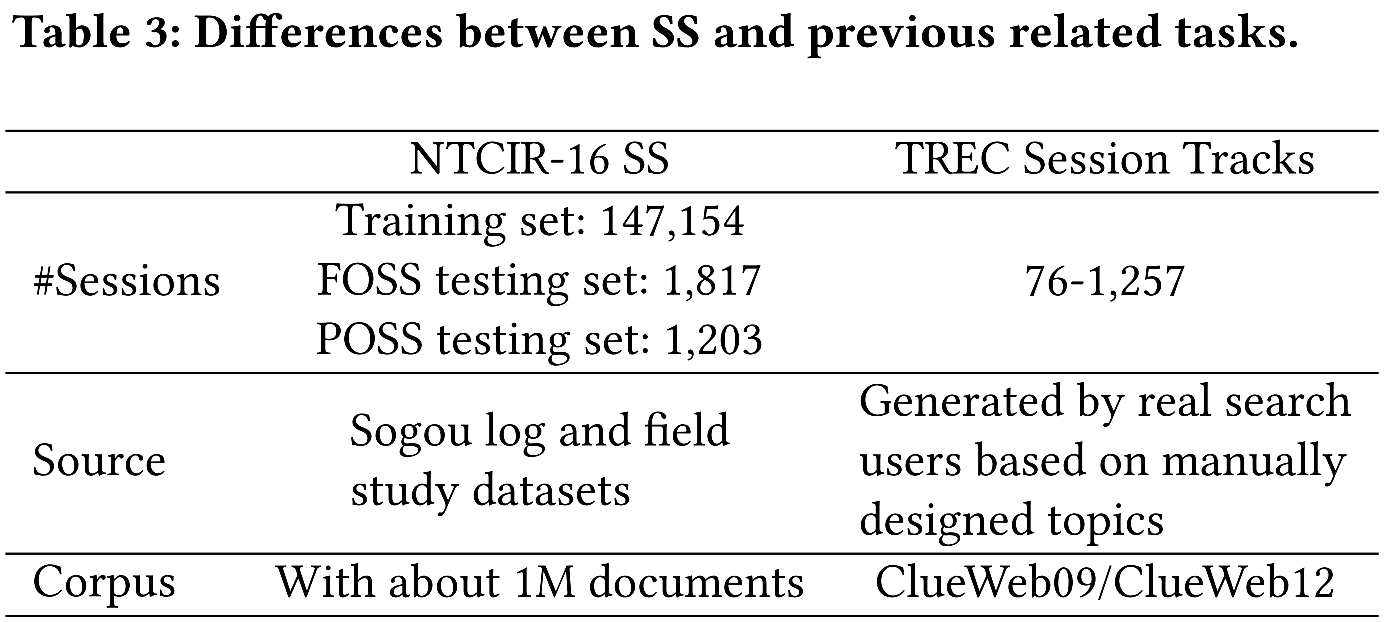


Table 3: Differences between SS and previous related tasks.

	NTCIR-16 SS	TREC Session Tracks
#Sessions	Training set: 147,154 FOSS testing set: 1,817 POSS testing set: 1,203	76-1,257
Source	Sogou log and field study datasets	Generated by real search users based on manually designed topics
Corpus	With about 1M documents	ClueWeb09/ClueWeb12

Dataset & Resources

```
SessionID       87
----------------------------
画杨桃    q198     1427848224.93
1       http://www.lbx777.com/yw06/x_hyt/kewen.htm      d1882   404     0       -1
2       http://pic.sogou.com/pics?query=%BB%AD%D1%EE%CC%D2&p=40230500&st=255&mode=255   d1883   <unk>   0       -1
3       http://tv.sogou.com/v?query=%BB%AD%D1%EE%CC%D2&p=40230600&tn=0&st=255   d1884   画杨桃-搜索页      0       -1
4       http://baike.sogou.com/v8080089.htm     d1885   画杨桃    0       -1
5       http://www.lspjy.com/thread-112497-1-1.html     d1886   人教版小学三年级下册语文《画杨桃》教学设计优质课教案      0       -1
6       http://weixin.qq.com/   d5      微信，是一个生活方式       0       -1
7       http://wenku.baidu.com/view/fa4903205901020207409c89.html       d1887   【图文】画杨桃_百度文库      0       -1
8       http://www.21cnjy.com/2/8135/   d1888   画杨桃课件_      0       -1
9       http://wenwen.sogou.com/s/?sp=S%E7%94%BB%E6%9D%A8%E6%A1%83      d1889   搜狗搜索  0       -1
10      http://www.aoshu.com/e/20090604/4b8bcabd28495.shtml     d1890   画杨桃_三年级语文下册课件_奥数网      0       -1
----------------------------
画杨桃ppt课件      q199    1427848230.2
1       http://wenku.baidu.com/view/bfe0c8edf8c75fbfc67db205.html       d1894   【图文】画杨桃ppt课件精品_百度文库      1       1427848232.105
2       http://www.1ppt.com/kejian/8846.html    d1895   《画杨桃》PPT课件  0       -1
3       http://www.1ppt.com/kejian/8851.html    d1896   《画杨桃》PPT课件6        0       -1
4       http://www.xiexingcun.com/xy6/HTML/53403.html   d1897   《画杨桃》公开课ppt课件（24页）-免费高速下载    0       -1
5       http://renjiaoban.21jiao.net/sanxia/huayangtao/ d1898   <unk>   0       -1
6       http://yuwen.chazidian.com/kejian108520/        d1899   画杨桃ppt课件下载  0       -1
7       http://www.docin.com/d-239045.html      d1900   <unk>   0       -1
8       http://www.glzy8.com/show/162484.html   d1901   11画杨桃PPT课件_管理资源吧  0       -1
9       http://www.xiexingcun.com/xy6/HTML/17470.html   d1902   《画杨桃》ppt课件-免费高速下载        0       -1
10      http://www.xiexingcun.com/xy6/HTML/61592.html   d1903   《画杨桃》ppt课件【13页】-免费高速下载        0       -1
```

Relevance Assessment

- Annotation company: *Panshidata.com* (盘石数据)
- Relevance assessment
  - 0 (irrelevant) – This HTML page is absolutely irrelevant to the user’s search intent, or it is a spam web page.
  - 1 (marginal relevant) – Users can obtain a small proportion of relevant information from the page
  - 2 (relevant) – This page contains most relevant information to user search intent
  - 3 (highly relevant) – This page should be ranked top among all pages, i.e., the navigational or official page of the query keyword

Relevance Assessment

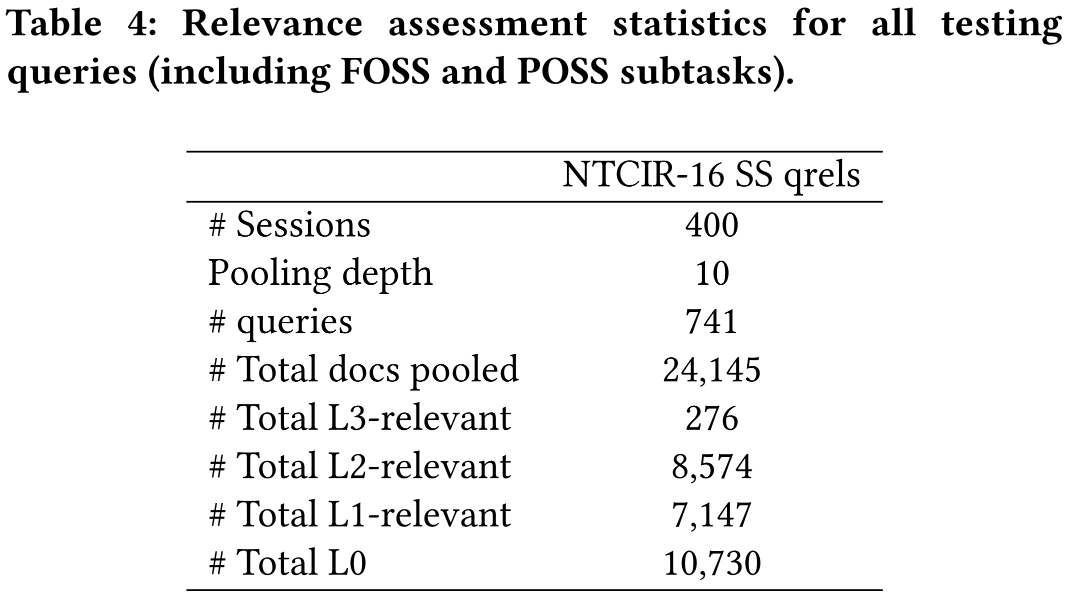


Table 4: Relevance assessment statistics for all testing queries (including FOSS and POSS subtasks).

	NTCIR-16 SS qrels
# Sessions	400
Pooling depth	10
# queries	741
# Total docs pooled	24,145
# Total L3-relevant	276
# Total L2-relevant	8,574
# Total L1-relevant	7,147
# Total L0	10,730

Evaluation Metrics

- FOSS

$$DCG@k = \sum_{i}^{K} \frac{2^{r(i)} - 1}{\log_2(i + 1)},$$

$$NDCG@k = \frac{DCG@k}{IDCG},$$

- POSS

$$RS - DCG = \sum_{m=1}^{M} mem_m \sum_{n=1}^{N} g(r_{m,n}, q_m) \cdot d_{m,n}(sDCG),$$

$$RS - RBP = \sum_{m=1}^{M} mem_m \sum_{n=1}^{N} g(r_{m,n}, q_m) \cdot d_{m,n}(sRBP),$$

$$mem_m = FF(M - m) = e^{-\lambda(M-m)}$$

$$d_{m,n}(sDCG) = \frac{1}{(1 + \log_{b_r} n)(1 + \log_{b_q} m)},$$

$$d_{m,n}(sRBP) = (\frac{p - bp}{1 - bp})^{m-1}(bp)^{n-1},$$

Evaluation Results (Preliminary)

Table 5: Preliminary evaluation results on the FOSS task (Sorted by the NDCG@3 score).

Rank	Run Name	nDCG@3	nDCG@5	nDCG@10
1	THUIR2-FOSS-NEW-5	0.0959	0.1185	0.1380
2	RUCIR-FOSS-REP-21	0.0406	0.0660	0.1043
3	RUCIR-FOSS-REP-31	0.0362	0.0677	0.1264
4	SCIR-FOSS-NEW-1	0.0356	0.0484	0.0609
5	RUCIR-FOSS-REP-3	0.0340	0.0429	0.0849
6	RUCIR-FOSS-REP-22	0.0293	0.0522	0.0864
7	THUIR2-FOSS-NEW-2	0.0225	0.0388	0.0833
8	RUCIR-FOSS-REP-2	0.0202	0.0276	0.0810
9	MM6-FOSS-REP-1	0.0183	0.0276	0.0311
10	THUIR2-FOSS-NEW-6	0.0169	0.0280	0.0683
11	MM6-FOSS-REP-3	0.0160	0.0203	0.0246
12	THUIR3-FOSS-REP-1	0.0156	0.0328	0.0556
13	MM6-FOSS-NEW-15	0.0152	0.0253	0.0414
14	THUIR2-FOSS-NEW-3	0.0130	0.0210	0.0626
15	MM6-FOSS-NEW-1	0.0128	0.0231	0.0419
16	THUIR1-FOSS-REP-1	0.0123	0.0327	0.0897
17	MM6-FOSS-NEW-2	0.0118	0.0200	0.0480
18	MM6-FOSS-REP-2	0.0115	0.0135	0.0170
19	RUCIR-FOSS-REP-1	0.0108	0.0248	0.0746
20	MM6-FOSS-NEW-21	0.0092	0.0176	0.0365
21	MM6-FOSS-REP-4	0.0081	0.0177	0.0428
22	MM6-FOSS-NEW-3	0.0075	0.0153	0.0311

Run Name	Description
MM6-POSS-REP-1	Hierarchical Behavior Aware Transformers
MM6-POSS-REP-2	HBA without history information
RUCIR-FOSS-REP-1	bert-base with ad-hoc fine-tune
RUCIR-FOSS-REP-2	BERT+Contrastive Learning
RUCIR-FOSS-REP-21	BERT+CL
RUCIR-FOSS-REP-22	BERT+CL+no-revise
RUCIR-FOSS-REP-3	BERT+CL+BM25
RUCIR-FOSS-REP-31	BERT+CL+BM25
RUCIR-POSS-REP-1	BERT+CL+BM25
RUCIR-POSS-REP-2	BERT+CL
RUCIR-POSS-REP-3	BERT+CL+BM25
SCIR-FOSS-NEW-1	First try with pyserini
THUIR1-FOSS-REP-1	naive bm25
THUIR2-FOSS-NEW-2	bm25 + tf-idf + f1-exp
THUIR2-FOSS-NEW-3	Bert with Ad-hoc Data Fine-tune
THUIR2-FOSS-NEW-4	Bert with Click Model Fine-tune
THUIR2-FOSS-NEW-5	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR2-FOSS-NEW-6	Bert with Session Data Fine-tune
THUIR2-POSS-NEW-1	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR3-FOSS-REP-1	THUIR3-BM25-test

Evaluation Results (Preliminary)

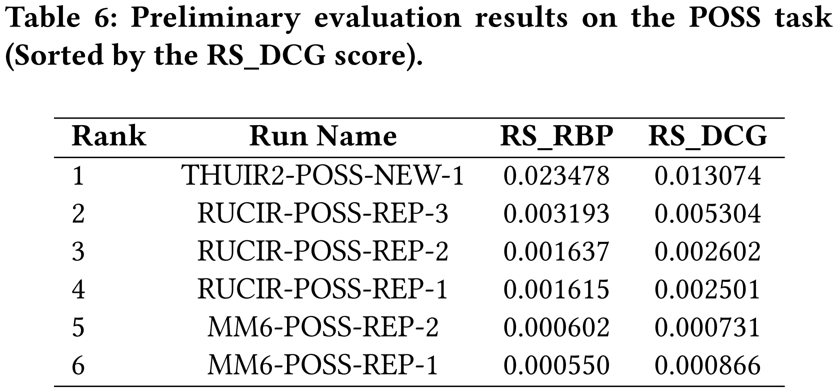
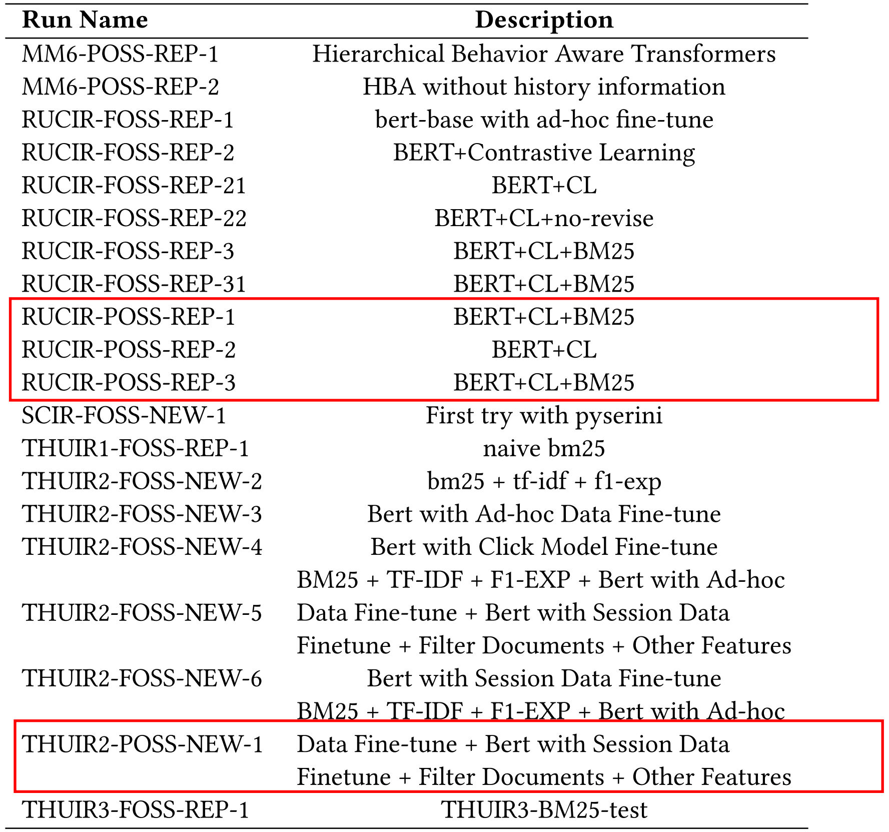


Table 6: Preliminary evaluation results on the POSS task (Sorted by the RS_DCG score).

Rank	Run Name	RS_RBP	RS_DCG
1	THUIR2-POSS-NEW-1	0.023478	0.013074
2	RUCIR-POSS-REP-3	0.003193	0.005304
3	RUCIR-POSS-REP-2	0.001637	0.002602
4	RUCIR-POSS-REP-1	0.001615	0.002501
5	MM6-POSS-REP-2	0.000602	0.000731
6	MM6-POSS-REP-1	0.000550	0.000866



Run Name	Description
MM6-POSS-REP-1	Hierarchical Behavior Aware Transformers
MM6-POSS-REP-2	HBA without history information
RUCIR-FOSS-REP-1	bert-base with ad-hoc fine-tune
RUCIR-FOSS-REP-2	BERT+Contrastive Learning
RUCIR-FOSS-REP-21	BERT+CL
RUCIR-FOSS-REP-22	BERT+CL+no-revise
RUCIR-FOSS-REP-3	BERT+CL+BM25
RUCIR-FOSS-REP-31	BERT+CL+BM25
RUCIR-POSS-REP-1	BERT+CL+BM25
RUCIR-POSS-REP-2	BERT+CL
RUCIR-POSS-REP-3	BERT+CL+BM25
SCIR-FOSS-NEW-1	First try with pyserini
THUIR1-FOSS-REP-1	naive bm25
THUIR2-FOSS-NEW-2	bm25 + tf-idf + f1-exp
THUIR2-FOSS-NEW-3	Bert with Ad-hoc Data Fine-tune
THUIR2-FOSS-NEW-4	Bert with Click Model Fine-tune
THUIR2-FOSS-NEW-5	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR2-FOSS-NEW-6	Bert with Session Data Fine-tune
THUIR2-POSS-NEW-1	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR3-FOSS-REP-1	THUIR3-BM25-test

Evaluation Results (Final)

Table 8: Final evaluation results on the FOSS task (Sorted by the NDCG@10 score).

Rank	Run Name	nDCG@3	nDCG@5	nDCG@10
1	RUCIR-FOSS-REP-31	0.5525	0.5623	0.5693
2	RUCIR-FOSS-REP-21	0.5365	0.5406	0.5570
3	THUIR2-FOSS-NEW-2	0.5229	0.5419	0.5525
4	RUCIR-FOSS-REP-3	0.4783	0.4785	0.4939
5	THUIR3-FOSS-REP-1	0.4805	0.4735	0.4636
6	MM6-FOSS-REP-1	0.4253	0.4420	0.4572
7	SCIR-FOSS-NEW-1	0.4620	0.4544	0.4309
8	THUIR2-FOSS-NEW-5	0.4068	0.4203	0.4156
9	MM6-FOSS-NEW-21	0.3642	0.3850	0.4029
10	MM6-FOSS-NEW-15	0.3587	0.3730	0.3925
11	THUIR2-FOSS-NEW-6	0.3431	0.3597	0.3684
12	THUIR1-FOSS-REP-1	0.1618	0.1785	0.1933

Run Name	Description
MM6-POSS-REP-1	Hierarchical Behavior Aware Transformers
MM6-POSS-REP-2	HBA without history information
RUCIR-FOSS-REP-1	bert-base with ad-hoc fine-tune
RUCIR-FOSS-REP-2	BERT+Contrastive Learning
RUCIR-FOSS-REP-21	BERT+CL
RUCIR-FOSS-REP-22	BERT+CL+no-revise
RUCIR-FOSS-REP-3	BERT+CL+BM25
RUCIR-FOSS-REP-31	BERT+CL+BM25
RUCIR-POSS-REP-1	BERT+CL+BM25
RUCIR-POSS-REP-2	BERT+CL
RUCIR-POSS-REP-3	BERT+CL+BM25
SCIR-FOSS-NEW-1	First try with pyserini
THUIR1-FOSS-REP-1	naive bm25
THUIR2-FOSS-NEW-2	bm25 + tf-idf + f1-exp
THUIR2-FOSS-NEW-3	Bert with Ad-hoc Data Fine-tune
THUIR2-FOSS-NEW-4	Bert with Click Model Fine-tune
THUIR2-FOSS-NEW-5	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR2-FOSS-NEW-6	Bert with Session Data Fine-tune
THUIR2-POSS-NEW-1	BM25 + TF-IDF + F1-EXP + Bert with Ad-hoc Data Fine-tune + Bert with Session Data Finetune + Filter Documents + Other Features
THUIR3-FOSS-REP-1	THUIR3-BM25-test

Evaluation Results (Final)

Table 9: Final evaluation results on the POSS task (Sorted by the RS_DCG score).

Rank	Run Name	RS_RBP	RS_DCG
1	RUCIR-POSS-REP-3	0.543875	0.746603
2	THUIR2-POSS-NEW-1	0.537723	0.648428
3	RUCIR-POSS-REP-1	0.473753	0.628068
4	RUCIR-POSS-REP-2	0.435542	0.563962
5	MM6-POSS-REP-2	0.326737	0.423102
6	MM6-POSS-REP-1	0.299660	0.379261

References

- [1] Ben Carterette, Paul Clough,Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The TREC session track 2011-2014. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 685–688.
- [2] Jia Chen, Jiaxin Mao,Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma.2021. Towards a better understanding of query reformulation behavior in web search. In Proceedings of the Web Conference 2021. 743–755.
- [3] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma.2019.TianGong- ST: A new dataset with large-scale refined real-world web search sessions. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2485–2488.
- [4] Kalervo Jarvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In European Conference on Information Retrieval. Springer, 4–15.
- [5] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, and Shaoping Ma. 2018. To- wards designing better session search evaluation metrics. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 1121–1124.

References

- [6] Tetsuya Sakai. 2013. Metrics, statistics, tests. In PROMISE winter school. Springer, 116–163.
- [7] TetsuyaSakai,SijieTao,ZhaohaoZeng,YukunZheng,JiaxinMao,ZhuminChu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, et al. 2020. Overview of the NTCIR-15 we want web with CENTRE (WWW-3) task. Proceedings of NTCIR-15. to appear (2020).
- [8] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview.. In TREC.
- [9] Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Cascade or recency: Constructing better evaluation metrics for session search. In Proceedings of the 43rd international acm sigir conference on research and development in information retrieval. 389–398.

THANK YOU!

We thanks all the NTCIR organizers and participants

for their sufficient support to our task!