Overview of the NTCIR-16 Session Search (SS) Task

Jia Chen
BNRist, Tsinghua University
Beijing, China
chenjia0831@gmail.com

Weihao Wu
Tsinghua University
Beijing, China
wuwh19@mails.tsinghua.edu.cn

Jiaxin Mao
Renmin University of China
Beijing, China
maojiaxin@gmail.com

Beining Wang
Tsinghua University
Beijing, China
wang-bn19@mails.tsinghua.edu.cn

Fan Zhang
Wuhan University
Wuhan, China
fan.zhang@whu.edu.cn

Yiqun Liu
BNRist, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

This is an overview of the NTCIR-16 Session Search (SS) task. The task features the Fully Observed Session Search subtask (FOSS) and the Partially Observed Session Search subtask (POSS). This year, we received 28 runs from 6 teams in total. This paper will describe the task background, data, subtasks, evaluation measures, and the evaluation results, respectively.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Users and interactive retrieval**;

KEYWORDS

session search, document ranking

1 INTRODUCTION

This paper presents an overview of the NTCIR-16 Session Search (SS) task [1]. SS is a pilot task in the NTCIR conference, aiming at exploring better ranking approaches for context-aware search scenarios. Nowadays, users depend increasingly on search engines to gain useful information or complete specific tasks. In complex search scenarios, a single query may not fully cover users' information needs. Therefore, search users will submit more queries to search systems within a short time interval until they give up the search process or their intents are satisfied. Such a process is called a search session or a search task. As user intent may evolve within search sessions, their actions and decisions will also be significantly impacted. Going beyond ad hoc search and considering the contextual information within sessions has been proved efficient for user intent modeling so far. However, existing tasks in NTCIR have not involved session-based retrieval yet. To this end, we propose this Session Search task to provide practical datasets and evaluation methodology to researchers in the related domain.

SS mainly consists of two subtasks: *Fully Observed Session Search* (FOSS) and *Partially Observed Session Search* (POSS). In the FOSS task, we provide full session contextual information within sessions to enhance the search effectiveness of the last query. While in the POSS task, we truncate all sessions before the last query. Participants need to leverage the limited search contexts to improve document ranking performance in the last multiple queries. In a nutshell, the FOSS task follows the basic idea in the TREC Session Tracks that the last query in a session is tend to be the most appropriate one that reflects user's information need. It is therefore crucial to improve user's search experience for this query. As for the POSS scenario, we suppose that sometimes users may not interact with the result page frequently. To this end, we only provide query sequence for the last queries within a session. This setting can also facilitate multi-query search effectiveness evaluation.

We provide a large-scale session dataset to support the training for various models. The training set was organized from a publicly available log-based session collection called TianGong-ST [3]. It contains about 150k refined web search sessions and human relevance labels for the last query of 2k sessions. For testing set, we extract search sessions from two field study datasets: TianGong-SS-FSD [9] and TianGong-Qref [2]. The two datasets contain abundant user interaction information such as click-through actions, query reformulating behaviors and explicit user feedback (instant relevance or usefulness annotations). Finally, there are 1,817 sessions in the FOSS subtask and 1,203 sessions in the POSS subtask.

Timeline of the NTCIR-16 Session Search task is shown in Table 1. This year we received 28 runs form six teams in total. The statistics are given in Table 2.

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

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)

[1] http://www.thuir.cn/session-search/

The remainder of this paper is organized as follows. Section 2 briefly introduces the background of related tasks and the motivation for setting up the SS task. Section 3 describes the details of the dataset processing and the relevance assessing process. In section 4, we list the subtask settings and the corresponding evaluation methodology. Section 5 reports the results of both preliminary and final evaluation. Finally, we conclude this paper in Section 6.

2 BACKGROUND

Related web search tasks such as Session Tracks [1] and Dynamic Domain (DD) Tracks [8] have disappeared from TREC for years. Among them, TREC Session Tracks (2011-2014) have been widely used as the benchmark for session search by numerous researchers. However, Session Tracks have some limitations: (1) They provided small-scale datasets (i.e., tens to thousands of web sessions) that can hardly support the training of more sophisticated models such as neural ranking models. (2) Based on simulated or manually designed search tasks, they collected session data via crowdsourcing experiments which can not reveal realistic web search scenarios. (3) They only evaluated the document ranking performance for the last query of a session by using ad hoc metrics such as Expected Reciprocal Rank (ERR), Normalized Discounted Cumulative Gain (NDCG), and Average Precision (AP) [6], etc. With the recent development of session-level evaluation metrics [5, 9], employing these metrics may yield more accurate assessments for the whole-session search effectiveness.

Besides Session Tracks, Dynamic Domain (DD) tracks (2015-2017) have also attracted some attention in investigating the interactive search process. Unlike Session Tracks, DD tracks focused more on the diversity of recalled documents and evaluated the systems with session-level metrics such as Cube Test, sDCG [4], and Expected Utility. Although they have achieved some success in studying one aspect of the interactive search process, there are limitations: (1) The assumption that their settings are based on is not that accurate. They preferred that an ideal system should automatically change their search strategies without users' query reformulating actions. However, query reformulation is a crucial behavior in realistic web search scenarios which explicitly represents users' evolving intents. Without query reformulation, it is hard for search engines to satisfy users' dynamic information needs. (2) The click-through data was generated by user simulators, while relevance annotations were collected from a third party. This setting could cause a mismatch between the relevance assessment and the simulated user behavior.

Based on these considerations, we would like to propose a new Session Search task in NTCIR-16. Compared to previous related tasks, we aim to create a more realistic and stable environment in which participants can evaluate session-based or task-oriented web search processes with both ad-hoc and session-level metrics. To support model training and testing, we will provide a large-scale practical session dataset refined from Sogou [2] search logs and several field study-based session collections. Our ultimate goal is to facilitate the investigation as well as the evaluation of complex search conditions. We believe that this task will benefit the community in various aspects, including system design, performance evaluation, user behavior prediction, and so on.

3 DATA COLLECTION

3.1 Session Data Preparation

For the training set, we use the TianGong-ST dataset [3]. It is a large-scale web search session dataset refined from an 18-day log of the Sogou search engine. TianGong-ST contains over 100k realistic search sessions and provides a subset of 2,000 sessions with human relevance labels. Participants can leverage the abundant click-through signals as well as the query reformulations within sessions to optimize their models.

We then extracted testing sessions from two field study datasets: TianGong-SS-FSD [9] and TianGong-Qref [2], both of which contain Chinese-centric search logs collected from tens of search users for over a month. To extract search sessions, we split the queries submitted by a user into sessions with a 30-minute gap. All sessions with at least two queries were considered valid sessions in our task. We then merged sessions extracted from two datasets as the testing set. As there is only one query before the last query in short sessions (length=2), these sessions will be directly used in the FOSS task. Then the remaining sessions (length > 2) will be randomly assigned to the FOSS and POSS subtasks with a ratio of about 1: 2. Finally, we obtained about 1.8k FOSS sessions and 1.2k POSS sessions. Figure ?? is an example of training session data format. We provided information such as query sequence, result URL, result title, click timestamps. As for testing sessions, we provided instant user usefulness annotations (extracted from the original field study datasets) so that participants could leverage these explicit feedbacks to optimize their training approaches.

As it may be difficult to handle HTML content, we provide a collection of about 1M documents with the preprocessed text content. For training queries, we directly used the corpus provided by TianGong-ST as the document collection. Then for all testing queries that need to return a re-ranked document list, we crawled the top 50 results from both *Baidu* and *Bing* search for each of them. As a result, there are about 80 candidate documents on average for all queries in the testing set. Participants need to re-rank these candidates according to given contextual information.

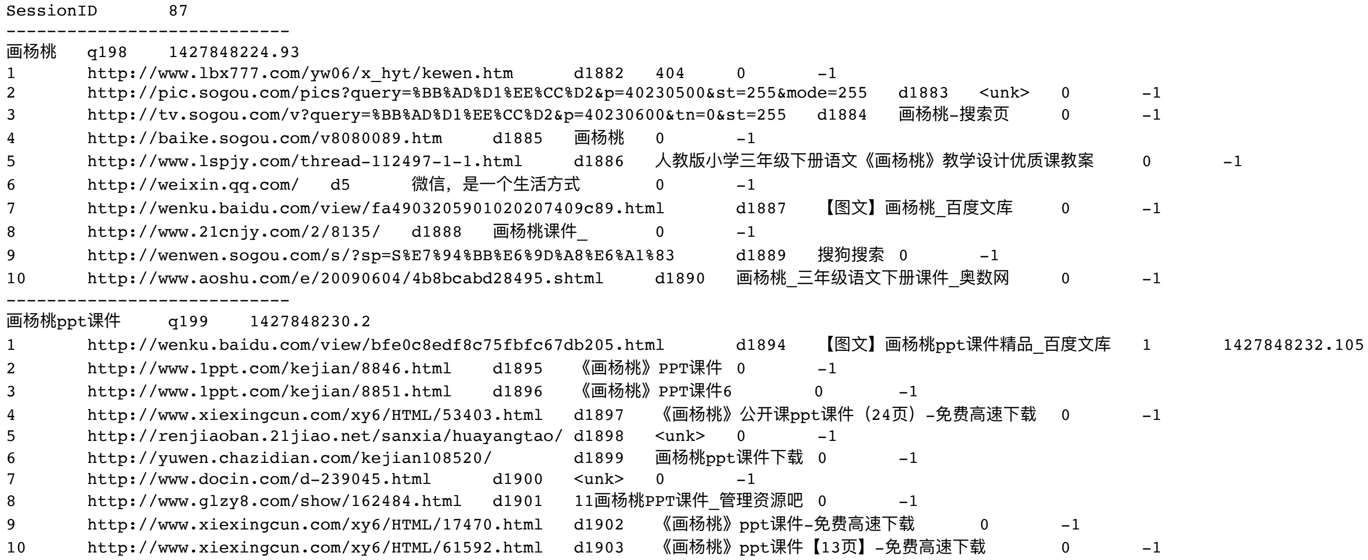
Table 3: Differences between SS and previous related tasks.

	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

3.2 Relevance Assessment

After the formal run process, all teams could select at most three best runs for each subtask for final system evaluation. We used a

[2] Sogou.com



```
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
```

Figure 1: Session data format in SS.

pooling depth of 10 (we only calculate the NDCG@10 scores) and randomly sampled 200 from two subtasks, respectively. In total, we needed to annotate 26,767 query-document pairs. We contacted a Chinese annotation company named 盘石数据 [3]. The relevance assessment lasted from February 20 to March 4, 2022. We provided query text, document text, and document URL to annotators for each query-document pair. They needed to assess the 4-scale relevance for each document according to the possible intents behind the query. Relevance assessment criteria are as follows:

- 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.

All the query-document pairs were annotated by three different experts. We used the median value as the final relevance label. Basic statistics of the relevance assessment are given in Table 4. We can observe that the number of irrelevant documents is the largest. L1- and L2-relevant documents account for about 65% of total pooled documents. This proportion is much larger than that of the previous relevance assessment [7], indicating the high quality of the top 50 documents of the two commercial search engines.

4 SUBTASKS AND EVALUATION METHODOLOGY

In this section, we briefly introduce the settings of our two subtasks: Fully Observed Session Search (FOSS) and Partially Observed Session Search (POSS).

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

4.1 Fully Observed Session Search (FOSS)

For a k-length session, we provide full session contexts in the first $(k-1)$ queries. Participants need to re-rank the candidate documents for the last query of a session. This setting follows TREC Session Tracks to enable ad-hoc evaluation by using metrics such as NDCG, AP, and RBP, etc. We choose NDCG@k as the evaluation metric in this subtask. Specifically, NDCG@K can be formalized as follows:

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

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

where $IDCG$ is the ideal discounted normalized gain based on all pooled documents of a query and $r(i)$ is the true relevance of the i-th document in the result list. Here we consider $k = 3, 5, 10$ in the FOSS subtask.

[3] https://www.panshidata.com

4.2 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 m queries, where $1 \leq m \leq k{-}1$. The value of n varies in different sessions. Participants will need to re-rank documents for the last $k{-}m$ queries (query) according to the partially observed contextual information in previous search rounds. Session-level metrics such as RS-DCG and RS-RBP [9] will be adopted for the evaluation of system effectiveness. RS-DCG and RS-RBP can be represented as follows:

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

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

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

where mem_m denotes users' memory of a query and is expressed as an exponentially decaying function and $g(r_{m,n}, q_m)$ maps the relevance or usefulness score of the n-th result in q_m to a gain. M and N is the number of queries within the session and the number of documents for a query, respectively. $d_{m,n}(sDCG)$ and $d_{m,n}(sRBP)$ represent the session-level discount functions, which can be formularized as following equations:

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

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

For sDCG, b_r and b_q are two logarithm base discounts for the ranking position and query position. As for sRBP, b and p are the balance and persistence parameters, respectively. In our evaluation process, we empirically set $b_r = 1.30$, $b_q = 1.3$, $b = 0.6$, $p = 0.8$ by following previous work [9].

5 EVALUATION RESULTS

5.1 Preliminary Evaluation

To provide all teams with instant feedback on their submitted run file before the final relevance assessment, we used the usefulness labels of the top 10 results in the two field study datasets to conduct the preliminary evaluation. The results are presented in Table 5 and Table 6, respectively. Table 7 presents the brief descriptions of each run file.

From Table 5, we find that using the naive BM25 algorithm can already yield a very good (rank 4, SCIR-FOSS-NEW-1). The best run is "THUIR2-FOSS-NEW-5", which considers various features to conduct learning-to-rank. They found that using document ID as a feature can greatly boost the preliminary evaluation performance. This may be because the usefulness annotations in the field study datasets are highly position-biased. A smaller document ID may indicate that the document is ranked higher on the original search engine result page. On another side, combining BERT with session-level click information and BM25 scores shows promising performance in RUCIR runs. This is consistent with some previous work that reports the effectiveness of combining dense retrieval models such as BERT with some traditional models. One obvious conclusion is that utilizing search history or session contexts is helpful for improving the system performance.

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

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

5.2 Final Evaluation

Preliminary evaluation only considered the relevance of the top 10 results for each query. Therefore, the system effectiveness evaluation may be greatly impacted by some biases, e.g., position bias and exposure bias. As we pooled the top 50 results of two commercial search engines, there may also be highly relevant results beyond the first ten documents. Therefore, we needed to pool the documents returned by all teams and collect the relevance label for all these documents. Based on the full relevance labels, we give final evaluation results in Table 8 and 9.

In the final evaluation, the RUCIR team achieves the best performance. As the ranking position of documents may help little

Table 7: Brief descriptions of some runs.

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

in the final evaluation, ranks of the THUIR2 team drop a lot compared to the preliminary evaluation, especially in the FOSS subtask (THUIR2-FOSS-NEW-5/6). Another reason is that we found that the THUIR2 team did not include all queries in the submission file. For the sake of fairness, we deem that the nDCG scores of the missing queries are 0. Therefore, the rank of the THUIR2 team varies a lot from the preliminary evaluation to the final evaluation. Even so, their approach that combines BM25 with TF-IDF and F1-EXP document filtering achieves a good performance (rank 3 in FOSS), better than many runs which involve complicated neural architectures such as MM6 runs. This is interesting because neural models such as transformers are expected to outperform traditional ones by a large margin. Surprisingly, only using BM25 (SCIR-FOSS-NEW-1) can achieve the fifth-best performance in the FOSS subtask on the nDCG@3 metric, indicating the effectiveness of traditional methods on re-ranking the most relevant results. By leveraging the HBA model, the MM6 team is ranked at the 6-th position among all runs.

As for the POSS subtask, the RUCIR team still achieves the best performance. One observation is that the "MM6-POSS-REP-2" run is better than the "MM6-POSS-REP-1" run. We expected an opposite result because the "MM6-POSS-REP-1" run using an approach considering the session history. This phenomenon is different from that in the FOSS subtask. We guess that the contextual information in the POSS subtask is not that sufficient for improving the query performance after multiple turns. Therefore, introducing too much search history of the previous queries distant from the current query may hurt the system performance to a certain extent.

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

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

6 CONCLUSIONS

This paper provided an overview of the NTCIR-16 Session Search task. As a pilot task, SS received 28 runs from six teams in total this year. Through the evaluation results, we find that 1) traditional models such as BM25 are still strong baselines compared to sophisticated neural models, 2) properly utilizing session context information can help improve the search effectiveness to a certain extent, 3) combining transformer-based methods with traditional approaches can yield promising performance in both FOSS and POSS subtasks. Besides, we find it possible to utilize advanced session-level metrics such as RS-RBP and RS-RBP in the evaluation of POSS tasks.

One possible action in the future Session Search task may be involving multilingual datasets. We may also go deeper into session-level evaluation, e.g., designing more tasks or utilizing more advanced metrics.

7 ACKNOWLEDGEMENTS

We thanks all participants and the NTCIR organizers for their sufficient support to our task. Our task is supported by the Natural Science Foundation of China (Grant No. 61732008, 61902209), Tsinghua University Guoqiang Research Institute, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

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 Järvelin, 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. Towards designing better session search evaluation metrics. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1121–1124.

[6] Tetsuya Sakai. 2013. Metrics, statistics, tests. In *PROMISE winter school*. Springer, 116–163.

[7] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, 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.