NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001324

Overview of the NTCIR-17 Session Search (SS-2) Task

Haitao Li
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
liht22@mails.tsinghua.edu.cn

Jia Chen
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
chenjia0831@gmail.com

Jiannan Wang
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
wchiennan@gmail.com

Weihang Su
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
swh22@mails.tsinghua.edu.cn

Qingyao Ai
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
aiqy@tsinghua.edu.cn

Xinyan Han
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
hanxinya20@mails.tsinghua.edu.cn

Beining Wang
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China
wang-bn19@mails.tsinghua.edu.cn

Yiqun Liu
DCST, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

This is an overview of the NTCIR-17 Session Search (SS-2) task. The task features the Fully Observed Session Search subtask (FOSS), the Partially Observed Session Search subtask (POSS) and the Session-level Search Effectiveness Estimation subtask(SSEE). This year, we received 16 runs from 2 teams in total. This paper will describe the task background, data, subtasks, evaluation measures, and the evaluation results, respectively.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**.

KEYWORDS

session search, retrieval, document ranking

1 INTRODUCTION

The Session Search (SS-2) task is a core task in NTCIR-17 to support intensive investigations of session search or task-oriented search. SS-2 task aims to explore better ranking approaches for context-aware search scenarios. Nowadays, search engines play an increasingly important role in daily life [6, 7]. In complex search scenarios, an individual query may not satisfy the user's information needs. Thus, users will submit more queries to the search system within short time intervals until they are satisfied or give up. Such a search process is known as a search session task [3]. In session search, the user's intent may change. This presents challenges to current ad-hoc search techniques.

To promote the development of session search techniques, we organize session search tasks at NTCIR. As the second year of organizing SS, we still employ settings that support not only (1) large-scale practical session datasets for model training but also (2) both ad-hoc and session-level evaluation this year. We would update the testing set by collecting data via an upcoming field study. Besides the aforementioned settings, we would also involve a new subtask for participants to design better session-level search effectiveness evaluation metrics.

Specifically, SS-2 mainly consists of three subtasks: *Fully Observed Session Search* (FOSS), *Partially Observed Session Search* (POSS) and *Session-level Search Effectiveness Estimation* (SSEE). Following the previous setting [3], the foss subtask provides the full context information before the last query and the poss subtask only provides limited contextual information. As for the new task, participants need to utilize user feedback to construct new session-level search effectiveness evaluation measures in SSEE. We believe that this will facilitate the development of the IR community in the related domain.

We provide the same large-scale session dataset to facilitate the training of various models. The dataset comprises approximately 150k meticulously curated web search sessions, accompanied by human relevance labels specifically assigned to the last query of 2k sessions. For testing set, we re-collected session data containing abundant user interaction information through field study. In contrast to previous setups, SS-2 does not provide candidate documents for each query, but instead provides large corpus. The corpus corresponding to the test session is T^2Ranking [9]. T^2Ranking contains 2M unique passages from real-world search engines. Participants need to retrieve and rerank the relevant documents in corpus for each query.

Timeline of the NTCIR-17 Session Search task is shown in Table 1. This year we received 16 runs form two teams in total. The statistics are given in Table 2.

The structure of the remaining sections in this paper is outlined as follows. In Section 2, we elaborate on the dataset processing details and the process of evaluating relevance. Section 3 presents the subtask settings along with their respective evaluation methodologies. The outcomes of final evaluations is presented in Section 4. Finally, we conclude this paper in Section 5.

Table 1: NTCIR-17 SS-2 timeline (time zone: AOE).

Session Dataset Release	June 30, 2023
Formal Run	July 1 - August 23, 2023
Relevance Assessment	August, 2023
Evaluation Results Release	September 1, 2023

Table 2: NTCIR-17 SS-2 run statistics.

Team	FOSS	POSS	SSEE	Total
THUIR	5	5	0	10
BITIR	2	2	2	6
Total	7	7	2	16

2 DATA

2.1 Training data

NTCIR-17 SS-2 reuses the training data of SS, i.e., the TianGong-ST [2] dataset. TianGong-ST is a substantial web search session dataset derived from an 18-day log of the Sogou search engine. It encompasses a vast collection of over 100,000 realistic search sessions, including a curated subset of 2,000 sessions labeled for human relevance. Participants are encouraged to harness the extensive click-through signals and query reformulations present within these sessions to enhance and optimize their models. Figure 1(a) shows an example of training data. We provided information such as query sequence, result URL, result title, click timestamps.

2.2 Testing data

We update the test set of SS-2 by field study. Specifically, we build an experimental platform to conduct the field study. The experimental platform consists of a browser plug-in and an annotation platform. The browser plug-in can record users' daily search behavior. The annotation platform is employed to collect feedback information submitted by users. From April 5 to May 13, 2023, there are 47 participants recruited for the study. Each participant is familiar with basic search engine usage. After participants signed the consent form and agreed to the data collection policy, we conducted training for each participant to ensure understanding of the experimental process. During the field study, participants can turn on the browser plugin, which keeps track of their daily search behavior. Participants can review previous queries through the annotation platform and divide them into different sessions.

We mix the dataset from the field study with sessions from two other Chinese-centric datasets TianGong-SS-FSD [10] and TianGong-Qref [1]. As there is only one query before the last query in short sessions (length=2), These sessions are randomly assigned to FOSS and SSEE tasks. Then the remaining sessions (length>2) are randomly assigned to three tasks. Finally, we obtained 1184 FOSS sessions, 976 POSS sessions and 1174 SSEE sessions. Figure 1(b) shows an example of testing data. We provide instant user usefulness annotations (extracted from the original field study datasets) and query satisfaction score.

Instead of providing candidate documents for each query, the SS-2 task provides large corpus T^{2}Ranking contains 2M unique documents from real search engines, which cover a wide range of topics and can fulfill the needs of different queries. Participants need to retrieve and rank relevant documents from T^{2}Ranking for each query.

Table 3: Relevance assessment statistics for all testing queries.

	NTCIR-17 SS-2 qrels
# Total docs pooles	58,752
# Total L3-relevant	208
# Total L2-relevant	1,956
# Total L1-relevant	7,840
# Total L0-relevant	48,748

2.3 Relevance Assessment

After the formal run process, all teams can select up to at most five runs for each subtask. We used a pooling depth of 5 (we only calculate the NDCG@5 scores). Finally we annotated 58,752 query-document pairs. We apply the same annotation criteria as NTCIR-16. Annotators need to assess 4-levels relevance for each document according to the possible intent behind the query. Specific relevance criteria are defined as follows:

- 0 (irrelevant) : The document does not provide useful information about this query topic, and the document is completely irrelevant to the query.
- 1 (marginal relevant) : The user can get a small portion of information from the document that is relevant to the query.
- 2 (relevant) : This document provides extensive information on the topic.
- 3 (highly relevant) : This document is dedicated to this query topic. It is very authoritative and comprehensive.

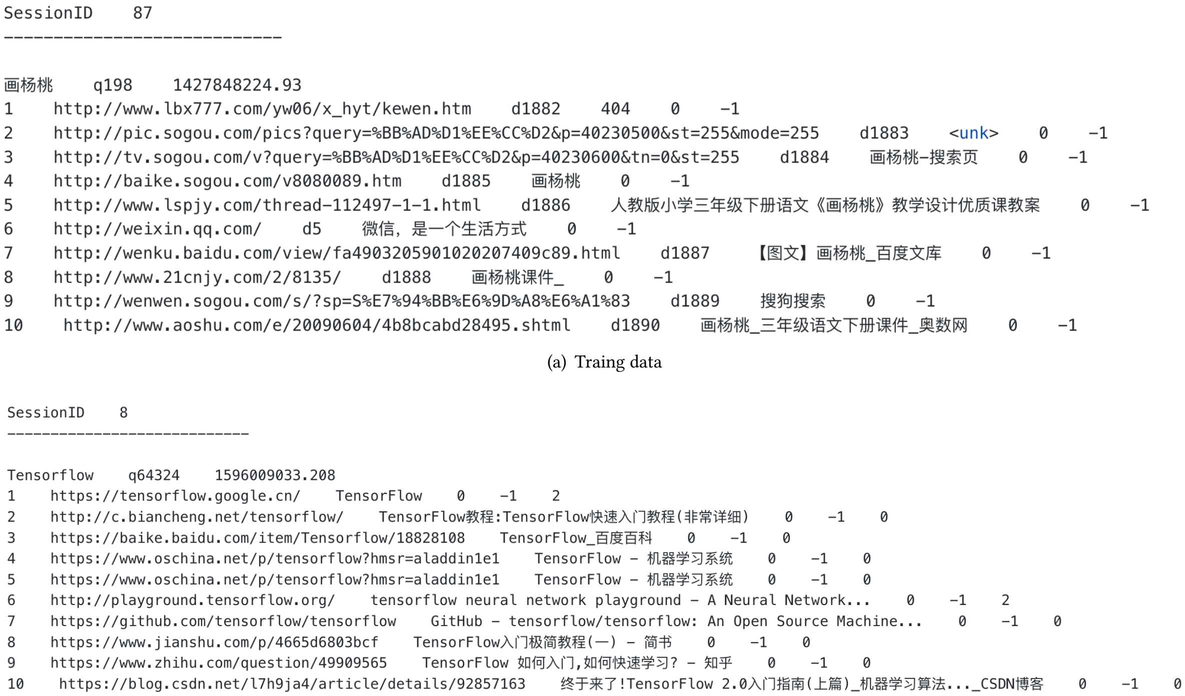
To ensure the quality of annotation, all query-document pairs are annotated by three different annotators. The final relevance label is the median value. Table 3 shows the statistical information of the relevance annotation. We can observe that the number of irrelevant documents is the largest. We assume that this is due to SS-2 requiring participants to retrieve the relevant documents by themselves, which is much more difficult. Thus we encourage participants to design better retrieval models [4, 5, 8].

3 SUBTASKS AND EVALUATION METHODOLOGY

In this section, we present the setting and evaluation criteria for the three subtasks: *Fully Observed Session Search* (FOSS), *Partially Observed Session Search* (POSS) and *Session-level Search Effectiveness Estimation* (SSEE). Table 4 summarizes the differences between SS-2 and previous related competitions.

3.1 Fully Observed Session Search (FOSS)

The FOSS subtask focuses on the effectiveness of the last query in the session, as it is often the most appropriate query to respond to the user's information needs. Formally, for a session with k queries, we provide full session contexts of top $k-1$ queries. Participants need to retrieve and rank the most relevant documents for the last



```
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/fa49032059010202074O9c89.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
```

(a) Traing data

```
SessionID    8
----------------------------

Tensorflow    q64324    1596009033.208
1    https://tensorflow.google.cn/    TensorFlow    0    -1    2
2    http://c.biancheng.net/tensorflow/    TensorFlow教程:TensorFlow快速入门教程(非常详细)    0    -1    0
3    https://baike.baidu.com/item/Tensorflow/18828108    TensorFlow_百度百科    0    -1    0
4    https://www.oschina.net/p/tensorflow?hmsr=aladdin1e1    TensorFlow - 机器学习系统    0    -1    0
5    https://www.oschina.net/p/tensorflow?hmsr=aladdin1e1    TensorFlow - 机器学习系统    0    -1    0
6    http://playground.tensorflow.org/    tensorflow neural network playground - A Neural Network...    0    -1    2
7    https://github.com/tensorflow/tensorflow    GitHub - tensorflow/tensorflow: An Open Source Machine...    0    -1    0
8    https://www.jianshu.com/p/4665d6803bcf    TensorFlow入门极简教程(一) - 简书    0    -1    0
9    https://www.zhihu.com/question/49909565    TensorFlow 如何入门,如何快速学习? - 知乎    0    -1    0
10    https://blog.csdn.net/l7h9ja4/article/details/92857163    终于来了!TensorFlow 2.0入门指南(上篇)_机器学习算法..._CSDN博客    0    -1    0
```

(b) Testing data

Figure 1: Session data format in SS-2.

Table 4: Differences between SS-2 and previous related tasks.

	NTCIR-17 SS-2	NTCIR-16 SS	TREC Session Track
#Sessions	Training set: 147,154 FOSS testing set: 1,184 POSS testing set: 976 SSEE testing set: 1,174	Training set: 147,154 FOSS testing set: 1,817 POSS testing set: 1,203	76-1,257
Source	TianGong-ST TianGong-SS-FSD/TianGong-Qref An un-released field study dataset	Sogou log and field study datasets	Generated by real search users based on manually designed topics
Document Collection	Training corpus: With about 1M documents. Test corpus: T2Ranking, with about 2.3M web pages.	With about 1M documents	ClueWeb09/ClueWeb12

query. FOSS subtask employs $NDCG@k$ as the evaluation metrics. The definition of $NDCG$ is as follows:

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

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

where $IDCG$ represents the ideal discounted normalized gain, calculated based on all aggregated documents of a query, and $r(i)$ signifies the true relevance of the i-th document in the result list.

3.2 Partially Observed Session Search (POSS)

The POSS subtask focus on last multiple queries search experience improvements. Formally, for a session with k queries ($k > 2$), We only provide session contexts for the first n queries ($1 \leq n \leq k-1$). The value of n varies in different sessions. Session-level metrics like RS-DCG and RS-RBP [9] will be utilized to assess the system's effectiveness. RS-DCG and RS-RBP can be formalized as follows:

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

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

$$\text{mem}_{\ m} = FF(M-m) = e^{-\lambda(M-m)}$$

A detailed description of the formula can be found at [10]. In our evaluation process, we employ the same hyperparameters as before.

3.3 Session-level Search Effectiveness Estimation (SSEE)

SSEE is a new subtask of SS-2 which aims to motivate participants to design better session-level search effectiveness evaluation metrics. We will provide a set of web sessions with full user interactive behaviors. Participants can utilize user feedback to construct new session-level search effectiveness evaluation measures. To meta-evaluate the reasonability of all proposed measures, we will compare the consistency of each measure and golden user satisfaction labels by calculating coefficients such as Pearson's γ and Spearman's ρ.

4 EVALUATION RESULTS

We pool all the documents returned by the run and collect the relevance labels for these documents. Based on all the relevance labels, we give the final evaluation results.

Table 5 shows the final results of the FOSS subtask. The THUIR team achieves the best performance. THUIR_SS-FOSS-NEW-3 combines lexical matching scores and semantic scores using a linear combination, which achieves the best results. THUIR_SS-FOSS-NEW-5 and THUIR_SS-FOSS-NEW-4 use the learning to rank algorithm to merge different features together. Their effects are also better than sparse or dense models alone. Surprisingly, the linear combination works better than the complex learning to rank algorithm. This may be due to the fact that there are changes in retrieval intent in many sessions, and these noises lead to a degradation in the performance of the learning to rank algorithm. In general, integrating features of different dimensions can help the model to achieve better results. How to consider the change of retrieval intent in session search is a potential future research interest.

As for POSS Task, the THUIR team still achieves the best results. It is worth noting that the performance of submission runs has decreased somewhat compared to NTCIR-16 SS. This reflects the increased difficulty of SS-2. Furthermore, all runs are not specifically designed for the characteristics of POSS subtask. Introducing too much search history of previous queries that are too far away from the current query may hurt system performance to some extent.

As for the SSEE subtask, it is a pity that only one team, BITIR, submits runs. BIT-SSEE-REP-2 achieves the best performance. It can be noticed that the session-level satisfaction predicted by the existing methods still differs a lot from the golden user satisfaction.

5 CONCLUSIONS

This paper provided an overview of the NTCIR-16 Session Search task. SS-2 received 16 runs from two teams in total this year. SS-2 updates the testing set by collecting data via field study and introduces a new subtask to design better session-level evaluation metrics. Through the evaluation, we found out that 1) Combining the scores of different features can effectively improve the performance of FOSS and POSS subtasks. 2) Session text information can improve performance to some extent, but they also have the potential to introduce noise. 3) Session-level satisfaction metrics need further refinement. In the future, we may consider multilingual datasets or design new subtasks to help better utilize session contexts.

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

Rank	Run Name	NDCG@3	NDCG@5
1	THUIR_SS-FOSS-NEW-3	0.5853154	0.6745773
2	THUIR_SS-FOSS-NEW-6	0.5643186	0.6569274
3	THUIR_SS-FOSS-NEW-5	0.3931865	0.4768206
4	THUIR_SS-FOSS-NEW-4	0.2506041	0.3309875
5	THUIR_SS-FOSS-NEW-1	0.1547940	0.2038491
6	BITIR-FOSS-NEW-2	0.0014880	0.0021785
7	BITIR-FOSS-NEW-1	0.0014391	0.0019561

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

Rank	Run Name	RS_DCG	RS_RBP
1	THUIR_SS-POSS-NEW-6	0.181201	0.379338
2	THUIR_SS-POSS-NEW-3	0.174898	0.367266
3	THUIR_SS-POSS-NEW-5	0.136628	0.288760
4	THUIR_SS-POSS-NEW-4	0.068510	0.143386
5	THUIR_SS-POSS-NEW-1	0.023533	0.048312
6	BITIR-POSS-NEW-1	0.000113	0.000250
7	BITIR-POSS-NEW-2	0.000028	0.000021

Table 7: Final evaluation results on the SSEE task (Sorted by the Pearson score).

Rank	Run Name	Pearson	Spearman
1	BITIR-SSEE-REP-2	0.4326276	0.4376210
2	BITIR-SSEE-REP-1	0.3878581	0.4076994

REFERENCES

[1] 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.

[2] 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.

[3] Jia Chen, Weihao Wu, Jiaxin Mao, Beining Wang, Fan Zhang, and Yiqun Liu. 2022. Overview of the NTCIR-16 session search (SS) task. *Proceedings of NTCIR-16. to appear* (2022).

[4] Qian Dong, Yiding Liu, Qingyao Ai, Haitao Li, Shuaiqiang Wang, Yiqun Liu, Dawei Yin, and Shaoping Ma. 2023. I3 Retriever: Incorporating Implicit Interaction in Pre-trained Language Models for Passage Retrieval. arXiv:2306.02371 [cs.IR]

[5] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. arXiv:2304.11370 [cs.IR]

[6] Haitao Li, Qingyao Ai, Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Zheng Liu, and Zhao Cao. 2023. Constructing Tree-based Index for Efficient and Effective Dense Retrieval. arXiv:2304.11943 [cs.IR]

[7] Haitao Li, Jia Chen, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Towards Better Web Search Performance: Pre-training, Fine-tuning and Learning to Rank. *arXiv preprint arXiv:2303.04710* (2023).

[8] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. arXiv:2305.06812 [cs.IR]

[9] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2Ranking: A large-scale Chinese Benchmark for Passage Ranking. *arXiv preprint arXiv:2304.03679* (2023).

[10] 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.