TianGong-ST: A New Dataset with Large-scale Refined Real-world Web Search Sessions†

Jia Chen, Jiaxin Mao, Yiqun Liu*, Min Zhang, Shaoping Ma
Department of Computer Science and Technology, Institute for Artificial Intelligence
Beijing National Research Center for Information Science and Technology
Tsinghua University, Beijing 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Web search session data is precious for a wide range of Information Retrieval (IR) tasks, such as session search, query suggestion, click through rate (CTR) prediction and so on. Numerous studies have shown the great potential of considering context information for search system optimization. The well-known TREC Session Tracks have enhanced the development in this domain to a great extent. However, they are mainly collected via user studies or crowdsourcing experiments and normally contain only tens to thousands sessions, which are deficient for the investigation with more sophisticated models. To tackle this obstacle, we present a new dataset that contains 147,155 refined web search sessions with both click-based and human-annotated relevance labels. The sessions are sampled from a huge search log thus can reflect real search scenarios. The proposed dataset can support a wide range of session-level or task-based IR studies. As an example, we test several interactive search models with both the PSCM and human relevance labels provided by this dataset and report the performance as a reference for future studies of session search.

CCS CONCEPTS

• **Information systems → Test collections**; *Web log analysis*; *Relevance assessment*;

KEYWORDS

Test collection; Session search; Information Retrieval

ACM Reference Format:
Jia Chen, Jiaxin Mao, Yiqun Liu*, Min Zhang, Shaoping Ma. 2019. TianGong-ST: A New Dataset with Large-scale Refined Real-world Web Search Sessions†. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3357384.3358158

1 INTRODUCTION

Considering session contexts is crucial for improving search systems. Take session search [8, 15] as an example: it aims to utilize the context information such as query sequences or user behaviors (clicks or scrolls) in previous search rounds to optimize the document ranking for following queries in the session. Moreover, numerous studies have shown great advantages of considering the context information for query suggestion [12]. Session contexts have also been taken into account in designing better session-level evaluation metrics [6] and predicting user satisfaction [4].

However, the lack of proper dataset limits the progress of related research. There are few test collections available for session-level IR research. Among them, TREC Session Tracks [2], running from 2011 to 2014, are the most widely applied datasets. They provide test collections with various forms of implicit feedbacks as well as human relevance labels for participants to optimize document ranking performance for the last query in a session. However, these tracks are mainly collected via user studies or crowdsourcing experiments with simulated search tasks. Therefore, they may not necessarily represent real-world Web search scenarios and only contain tens to thousands sessions that are usually deficient for more sophisticated models. TREC Dynamic Domain (DD) Tracks [14] provide both topic- and subtopic-level relevance annotations. Whereas, they adopt simulators to generate user feedbacks and focus merely on specific domains. Besides, the large-scale AOL search log [1] is collected from real users, but it is noisy and outdated (a certain proportion of URLs are no longer accessible). Ultimately, further studies in the domain call for larger-scale and more authentic benchmarks.

In this paper, we present a new Chinese-centric session dataset named TianGong-ST [1]. Containing 147,155 refined search sessions and 40,596 unique queries in total, the dataset is extracted from an 18-day search log from *Sogou*. We then apply six popular click models (TACM [7], PSCM [10], THCM [13], UBM [4], DBN [3] and POM [11]) to obtain unbiased relevance labels for each query-document pair. We also sample a subset of 2,000 sessions from TianGong-ST and collect session-level human relevance labels for documents of last queries in them. To show that this new dataset can facilitate the training and evaluation of session-based retrieval models, we further test a wide range of existing interactive search algorithms on it and present the results as references.

To summarize, TianGong-ST has the following advantages:

- Instead of relying on crowdsourcing or user studies, the sessions are sampled from a real, large-scale search engine log and refined through a series of processing steps. Our dataset can reflect more realistic Web search scenarios.
- It contains over a hundred thousand sessions with abundant click information as well as textual information including queries, titles and full-text documents.
- It can be used in a wide range of researches such as session search, query suggestion, CTR prediction, and etc. To demonstrate the effectiveness of the dataset in session search, we reproduce several typical interactive search algorithms and test their performances respectively. We then enclose the evaluation results based on both PSCM and human relevance labels for reference.

*Corresponding author

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358158

†This work is supported by the National Key Research and Development Program of China (2018YFC0831700) and Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011).
[1]This dataset is now available at http://www.thuir.cn/tiangong-st/

2 DATASET

2.1 Data Preparation

Our dataset is based on an 18-day (*April 1st - 18th, 2015*) span query log collected by a commercial search engine *Sogou.com*. For each search round, the query, result URLs with their corresponding vertical types and click information (whether be clicked and the click timestamps) have been recorded. The raw data contains abundant Web search sessions mingled with noise. Therefore, it is hard to directly employ it for research purpose. To tackle the issue, we refine the sessions through a series of procedures and filter the noisy data step by step. Table 1 shows the procedures in detail.

Table 1: Session Refining Procedures

#1	Split the queries into sessions by a 30-minute gap.
#2	Select sessions with lengths in range of 2-10.
#3	Filter sessions with semantic similarities between their last queries and previous ones less than 0.5.
#4	Remove sessions with queries whose frequency < 10.
#5	Select sessions with at least one click.
#6	Filter sessions which contain pornographic, violent or politically sensitive contents.
#7	Truncate documents for each query with a cut-off at 10.
#8	After crawling Web pages, remove sessions with over 20% missing documents.
#9	After supplementing some documents via Sogou-QCL, remove sessions with more than three missing documents.
#10	Filter sessions longer than three but with invariant queries.

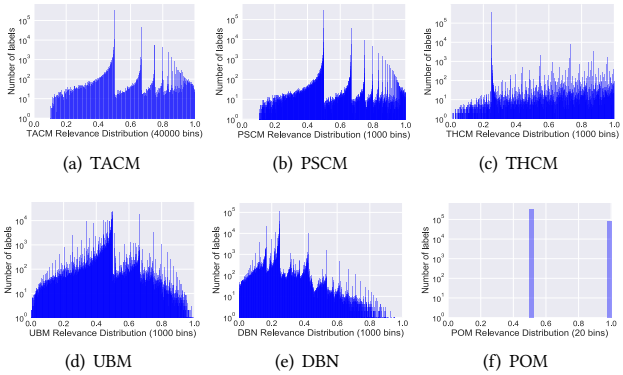
First of all, we adopt the widely-used 30-minute gap to split the queries into search sessions. To utilize context information, we exclude sessions with only one query. In addition, sessions that are too long, i.e. longer than ten queries, usually contain more noise but only account for a little proportion ($\leq 0.05\%$), hence have also been filtered. An investigation shows that a proportion of sessions are internally inconsistent. In other words, queries within a session possibly belong to very different topics. This may cause problems for follow-up studies. To resolve this situation, we deploy the open-source tool GloVe [9] to train word vectors on a large corpus from Sogou-QCL [16] and use the cosine similarity between the max-pooled GloVe vectors to measure the semantic similarities between queries. We filter out a session if the cosine similarity between its last queries and previous ones is less than 0.5 in Step 3. Here we search the threshold value in [0, 1] with the step of 0.01 and find 0.5 is the best value for the trade-off of ensuring session-level consistency and avoiding discarding too many good sessions.

In Step 4, we remove sessions containing rare queries that appear less than 10 times to protect user privacy. Only sessions with at least one click are retained in Step 5, because it is hard to make use of non-click sessions in terms of collecting user feedbacks, predicting click-based satisfaction, or evaluating system performances. Then in Step 6, we filter sessions containing pornographic, violent, or politically sensitive contents by using a huge sensitive word dictionary with the size of about 65,000. To check the effect of this step, we sample several subsets of 2,000 sessions and fail to find any sensitive contents. In Step 7, the document list for each query is truncated at rank 10 to unify the maximum position for click models. We further deploy additional measures to polish our dataset. To ensure the freshness of corpus, we crawl the Web pages for the whole dataset recently and remove sessions with over 20% missing documents. In the next, we supplement some documents for our corpus via Sogou-QCL and then discard sessions with more than three missing documents. Finally, we filter out sessions that are longer than three but contain a single repeated query.

Based on the remaining sessions, we train six click models (TACM, PSCM, THCM, UBM, DBN and POM) to obtain the click-based relevance labels. The average perplexities of different click models are shown in Table 2. Among these models, PSCM performs the best, followed by TACM. This finding is slightly different from the results in [7], where TACM performs better than PSCM. We further show the distributions of click-based relevance labels generated by different click models in Figure 1. From this figure, we can see that TACM and PSCM share similar, denser label distributions while POM only estimates very sparse scores. According to the perplexity, we choose PSCM to test system performances later.

Table 2: Average Perplexities for Various Click Models

TACM	PSCM	THCM	UBM	DBN	POM
1.0318	**1.0153**	1.3272	1.1867	1.1848	1.4653



(a) TACM (b) PSCM (c) THCM

(d) UBM (e) DBN (f) POM

Figure 1: Label distributions for six click models.

2.2 Overview of TianGong-ST

In this section, we will briefly introduce our dataset. The session data is organized in a prettified XML format similar to TREC Session Tracks. A session consists of several search interactions together with a clicked-document list. Each interaction represents a search iteration where a user submits an independent query and receives top 10 documents from the search engine. For each round of interactions, the query text and query identifier are provided. For each document in the result list, the URL, document identifier, title, and six click-based relevance labels are given. In addition, the start timestamps for all sessions, interactions, and clicked documents are also presented to support dwell-time based models. Titles of Web pages that we fail to crawl are replaced by UNK.

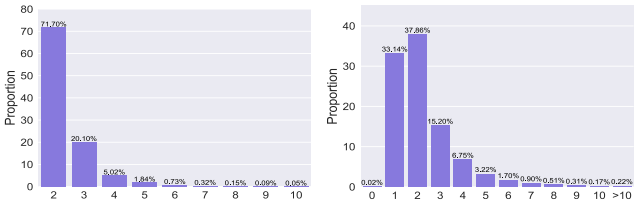
Scale of data. Table 3 represents the comparison between TianGong-ST and TREC Session Tracks. The released dataset compromises 147,155 full sessions, with 40,596 unique queries in total. For researchers' convenience, we provide a preprocessed corpus that covers over 90% Web pages (279,597 out of 309,287) involved in TianGong-ST. For other documents we fail to crawl the contents, we provide their URLs and click labels. We adopt an open source tool named `jieba_fast`[2] for Chinese word segmentation and release the preprocessed corpus. The documents in our corpus have an average length of 3269.75 characters (without word segmentation).

[2]https://pypi.org/project/jieba-fast/0.42/

Table 3: Comparison between TianGong-ST and TREC Session Track 2011-2014 [2]

Dataset	TREC 2011	TREC 2012	TREC 2013	TREC 2014	TianGong-ST
#sessions	76	98	133	1,257	147,155
#unique queries	280	297	442	3,213	40,596
#avg. session length	3.68	3.03	5.08	4.33	2.42
#avg. click per session	2.4	2.8	4.4	1.34	2.25
#relevance judgments	19,413	17,861	13,132	16,949	20,000
search engine	BOSS+CW09 filter	BOSS+CW09 filter	indri	indri	*Sogou.com*
collection	ClueWeb09	ClueWeb09	ClueWeb12	ClueWeb12	Newly crawled in December, 2018

Session lengths & clicks. As shown in Figure 2(a), over 70% sessions contain two queries, indicating that in real-world Web search environment users tend to submit a single query reformulation. An overwhelming majority of sessions fall into the 2-5 length interval. Figure 2(b) presents the click distribution of our datasets. Sessions with two clicked documents dominate with the largest proportion of 37.86%. Clicked documents normally serve as implicit user feedbacks and play an essential role in interactive systems. Copious click information in our dataset facilitates the investigation with more sophisticated models.



(a) Session length Distribution (b) Session Click Distribution

Figure 2: Distributions for session lengths and clicks.

Query reformulation. Apart from clickthrough data, query reformulation is another form of user feedback signals. Depending on the results of preceding queries and current information needs, users determine the following queries to submit. Thus query reformulations may imply users' intent shifts. To illuminate the composition of query reformulation types in TianGong-ST, we count the proportions of *Add*, *Delete*, *Change* and *Keep* types for all pairs of two consecutive queries, and compare them with those of a raw data sample (See Figure 3(a)). The four reformulation types are represented as follows ($+\triangle q_t/-\triangle q_t$ are defined according to [15]):

Add : $+\triangle q_t \neq \emptyset, -\triangle q_t = \emptyset$; **Delete** : $+\triangle q_t = \emptyset, -\triangle q_t \neq \emptyset$;

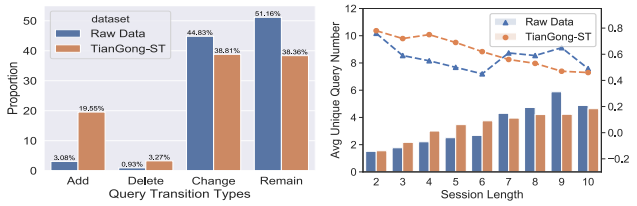
Change : $+\triangle q_t \neq \emptyset, -\triangle q_t \neq \emptyset$; **Keep** : $+\triangle q_t = \emptyset, -\triangle q_t = \emptyset$.

As shown in Figure 3(a), the *Change* type takes the greatest part among the four reformulation types in TianGong-ST, closely followed by *Keep*. The proportion of *Add* and *Delete* exceeds 22%, much more than that of the raw data (about only 4% in total). This indicates that our session refinement has balanced the proportions across all reformulation types. Table 4 presents some example sessions. In the first case, there is a specification then a parallel shift in the query reformulation process. While in the second case, the user firstly generalized the query from "Transformers 4" to "Transformers", then specified her intent on "Ultra Magnus", and ended the search process with "The Fallen". User may adopt different query reformulation strategies in different search tasks.

Unique query number. If there are excessive duplicated queries within a session, query-change based models may not work well. Figure 3(b) represents a comparison analysis for average unique query number across sessions with different lengths in TianGong-ST and raw data. We can find that, TianGong-ST owns more unique queries per session within the length interval of 2-6 where over 99% sessions fall into this region.

Table 4: Example query sequences within a session in the TianGong-ST (translated from Chinese).

Conan → *Conan* Movie Version→ *Conan* Mandarin Version
Transformers 4 → *Transformers* → *Ultra Magnus* → *The Fallen*
Minecraft → *Minecraft* skin websites
Tiantian Express → Postal Registered Letter Inquiry
the cause of high diastolic blood pressure and its harm → what to eat to lower diastolic blood pressure?



(a) Query Transition (b) Unique Query Number

Figure 3: Some statistics of TianGong-ST and the raw data.

2.3 Session-level Relevance Annotation

Click models usually estimate click-based relevance labels through various unbiased algorithms. However, sometimes a clicked document may not sufficiently be very relevant. Additionally, it is really expensive to collect human relevance labels for all query-document pairs. Therefore, we make a stratified sample of 2,000 sessions from TianGong-ST for session-level relevance annotation. To balance the number of sessions with different lengths, the distribution rates are 50%, 18%, 14%, 8%, 5%, 2%, 1%, 1%, 1% for lengths varying from 2 to 10 respectively after sampling.

We recruit 20 participants aged from 18 to 26 to annotate session-level relevances for documents in the last interaction of a session. All of the participants are familiar with the basic operations of Web search. Each participant completes annotating for 300 sessions and receives a reward of about $60. Due to substantial volumes of tasks, all participants have received guidance at the laboratory and are allowed to finish the tasks anywhere online. For each task, context information including the query sequence and clicked documents in previous search rounds within the session are presented to the participants. They should consider users' whole-session information needs accordingly and annotate for 10 results of the last query. To ensure the quality of annotation, the submit button of each session is enabled only if the annotator read all of the 10 documents for at least five seconds. As instructed, the annotator should infer not only users' intrinsic information needs from clicked documents, but also their shifted intents from query reformulations. According to

Table 5: The performances of some models on Test-PSCM and Test-HL (Human Label). The 95% confidence intervals of Gaussian distributions for all results are presented in the subscript. Note that * indicates a statistical significance over BM25 at p < 0.001 level.

Model	Test-PSCM				Test-HL			
	nDCG@1	nDCG@3	nDCG@5	RBP(0.8)	nDCG@1	nDCG@3	nDCG@5	RBP(0.8)
BM25	$0.4963_{\pm 0.0008}$	$0.5597_{\pm 0.0005}$	$0.6217_{\pm 0.0004}$	$0.4300_{\pm 0.0000}$	$0.4820_{\pm 0.0082}$	$0.5547_{\pm 0.0061}$	$0.6167_{\pm 0.0048}$	$0.2587_{\pm 0.0019}$
QCM SAT	$0.4969_{\pm 0.0008}$	$0.5506^{*}_{\pm 0.0005}$	$0.6105^{*}_{\pm 0.0004}$	$0.4287^{*}_{\pm 0.0003}$	$0.2622^{*}_{\pm 0.0077}$	$0.3837^{*}_{\pm 0.0061}$	$0.4657^{*}_{\pm 0.0049}$	$0.2332^{*}_{\pm 0.0020}$
Rocchio	$0.5413^{*}_{\pm 0.0008}$	$0.5916^{*}_{\pm 0.0008}$	$0.6465^{*}_{\pm 0.0007}$	$0.4326^{*}_{\pm 0.0007}$	$0.7197^{*}_{\pm 0.0085}$	$0.7050^{*}_{\pm 0.0056}$	$0.7379^{*}_{\pm 0.0046}$	$0.2832^{*}_{\pm 0.0022}$
Rocchio CLK	$\mathbf{0.5433}^{*}_{\pm 0.0008}$	$0.5930^{*}_{\pm 0.0008}$	$0.6474^{*}_{\pm 0.0007}$	$0.4327^{*}_{\pm 0.0007}$	$\mathbf{0.7288}^{*}_{\pm 0.0084}$	$0.7099^{*}_{\pm 0.0055}$	$0.7402^{*}_{\pm 0.0045}$	$0.2837^{*}_{\pm 0.0022}$
Rocchio SAT	$0.5428^{*}_{\pm 0.0008}$	$0.5929^{*}_{\pm 0.0008}$	$0.6472^{*}_{\pm 0.0007}$	$0.4327^{*}_{\pm 0.0007}$	$0.7282^{*}_{\pm 0.0084}$	$\mathbf{0.7102}^{*}_{\pm 0.0055}$	$\mathbf{0.7403}^{*}_{\pm 0.0045}$	$\mathbf{0.2837}^{*}_{\pm 0.0022}$
Win-win	$0.4781^{*}_{\pm 0.0007}$	$\mathbf{0.5968}^{*}_{\pm 0.0005}$	$\mathbf{0.6823}^{*}_{\pm 0.0004}$	$\mathbf{0.4334}^{*}_{\pm 0.0000}$	$0.4787_{\pm 0.0082}$	$0.5526_{\pm 0.0060}$	$0.6154_{\pm 0.0049}$	$0.2590_{\pm 0.0019}$

the annotation records, average 1.048 previously clicked documents have been checked by the participants per session, suggesting that they do consider the session contexts during the annotation.

We adopt a five-graded relevance label similar to TREC Session Tracks. Judgment values are: 0 for not relevant or spam, 1 for relevant, 2 for highly relevant, 3 for key, and 4 for navigational. Each query-document pair receives three labels from different annotators. We use the median of the three annotations as the final relevance label. A consistency check shows that the annotations achieve a weighted Kappa(κ) of 0.4826 (std=0.0025), which indicates a moderate consistency. Here we calculate the linear weighted Kappa instead of Fleiss's Kappa to count disagreements differently.

3 APPLICATION

Our dataset can be applied for multiple information retrieval tasks such as session search, query suggestion, click prediction, session-level relevance estimation and so on. Here we take session search as an example and test the performances of some systems based on PSCM and human labels. The baselines include BM25, QCM SAT [15], Rocchio [5], Rocchio CLK, Rocchio SAT, and Win-win model [8]. Some of these models are not open-sourced, so we implement them according to the published papers. We have incorporated some changes in these models to adapt to our dataset. For win-win model, the changes are: only BM25 algorithm is used as the core search strategy, the search engine actions only include changing the term weights, and user actions are [*Add*, *Delete*, *Keep*, *Change*]. We adopt five-fold cross-validation for win-win model, using the training set to estimate the parameters and learn the initial Q-learning action values. Then for both Test-PSCM and Test-HL, we run the pre-trained win-win model on the test set where new transitions of user decision states still occur. Note that the scales of labels in Test-PSCM and Test-HL are $[0, 1]$ and $\{0, 1, 2, 3, 4\}$, respectively.

Table 5 shows the performances of different ranking models. We only calculate the metrics with a cut-off at 5 because there are only 10 candidates for each query. We can observe that the win-win model achieves the best overall performance on Test-PSCM, which is the same as the results reported in [8]. However, it shows almost no advantage on Test-HL (very close to BM25). This may imply the win-win model needs more appropriate reward signals to learn a better Q-table. We can only take PSCM labels to formulate rewards for previous search rounds within a session but evaluate with human labels so the performances may drop. For Rocchio algorithms with three kinds of feedbacks, the differences are not obvious in both test conditions. It is weird that QCM SAT performs worst among all the models. Maybe it is sensitive to the system parameters but we directly use the empirical parameters reported in the original paper without tuning. Generally, feedback based models like Rocchio are usually more robust than query change based models (QCM, Win-win). This experiment demonstrates that TianGong-ST can effectively support the training and evaluation of different sessions search models. The results can further be used as baselines for novel session-based retrieval models.

4 DISCUSSIONS AND CONCLUSIONS

In this paper, we introduce a new dataset named TianGong-ST. It contains a large amount of refined Web search sessions with abundant click information as well as query reformulations. A corpus with high document coverage is also provided for researchers' convenience. Both click-based labels estimated by six popular click models for all query-document pairs and human relevance labels for 2,000 sampled sessions are available. To explore some features of it, we also conduct a detailed investigation of it. We further reproduce some baseline systems and report their performances on TianGong-ST to provide references for other researchers. Experiment results show that our dataset can be easily employed for session search task. There are some other usages for TianGong-ST such as query suggestion, CTR prediction, session-level relevance estimation, and etc. We hope that our dataset could provide more opportunities for researchers who are interested in relative domains.

REFERENCES

[1] David J Brenes and Daniel Gayo-Avello. 2009. Stratified analysis of AOL query log. *Information Sciences* 179, 12 (2009), 1844–1858.

[2] 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*. ACM, 685–688.

[3] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. *the web conference* (2009), 1–10.

[4] Georges Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. (2008), 331–338.

[5] Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science.

[6] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, and Shaoping Ma. 2018. Towards Designing Better Session Search Evaluation Metrics.. In *SIGIR*. 1121–1124.

[7] Yiqun Liu, Xiaohui Xie, Chao Wang, Jianyun Nie, Min Zhang, and Shaoping Ma. 2017. Time-Aware Click Model. *ACM Transactions on Information Systems* 35, 3 (2017), 16.

[8] Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: dual-agent stochastic game in session search. (2014), 587–596.

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. (2014), 1532–1543.

[10] Chao Wang, Yiqun Liu, Meng Wang, Ke Zhou, Jianyun Nie, and Shaoping Ma. 2015. Incorporating Non-sequential Behavior into Click Models. (2015), 283–292.

[11] Kuansan Wang, Nikolas Gloy, and Xiaolong Li. 2010. Inferring search behaviors using partially observable Markov (POM) model. (2010), 211–220.

[12] Bin Wu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Query Suggestion With Feedback Memory Network. *the web conference* (2018), 1563–1571.

[13] Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Incorporating revisiting behaviors into click models. (2012), 303–312.

[14] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview. In *TREC*.

[15] Sicong Zhang, Dongyi Guan, and Hui Yang. 2013. Query change as relevance feedback in session search. (2013), 821–824.

[16] Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-QCL: A New Dataset with Click Relevance Label. (2018), 1117–1120.