Data Description of SIGIR17 paper "Meta-evaluation of Online and Offline Web Search Evaluation Metrics"

Ye Chen
Tsinghua University
Beijing, China
chenye617@gmail.com

ABSTRACT

This paper provides a description of the datasets used in the SIGIR-17 paper "Meta-evaluation of Online and Offline Web Search Evaluation Metrics", including the organization of search tasks and ranking mechanisms of SERPs.

KEYWORDS

Data description, task organization, ranking mechanism

1 INTRODUCTION

The datasets used in the SIGIR17 paper "Meta-evaluation of Online and Offline Web Search Evaluation Metrics"[2] are composed of two parts, Dataset #1 and Dataset #2. These two datasets contain 2685 search sessions collected under 65 search search tasks in total. Some statistics of these two datasets are shown in Table 1.

Table 1: Characteristics of Datasets

	# queries	# different rankings per query	# users	# sessions
Dataset #1	26	3	40	1038
Dataset #2	30	6~10	98	1397

Both datasets contain the following information for each search session:

- Query and corresponding task descriptions.
- Information of ranked search results as shown on SERPs.
- 4-scaled relevance assessments of all search results.
- 5-scaled user satisfaction annotations.
- Users' interaction behaviors during the search process, including click-through, mouse hover and dwell time information.

With the rich information provided by the datasets, we meta-evaluate different evaluation metrics in the SIGIR17 paper. In this paper, we describe the data generation process in detail, as well as the organization of search tasks and ranking mechanisms of search engine result pages (SERPs).

2 DATA DESCRIPTION

2.1 Experiment Procedure

These two datasets are generated under the same experimental process which is shown in Figure 1. Each participant completed a series of no more than 30 tasks in the datasets and they were required to perform two warm-up search tasks first to get familiar with the search process. Before each task, the participant was shown the search query and task explanations (see the card on the top right corner of Figure 1 as an example) first to avoid ambiguity. After that, the participant would be guided to search result page where the query is not allowed to change. Each participant was asked to examine the 10 fixed search results provided by the system and end the search session either if the search goal was completed or he/she was disappointed with the results. During the search process, users' interaction behaviors, including mouse movements, clicks, hovers and scrolls are logged by injected Javascript on SERPs. The provided search result lists were pre-crawled from commercial search engines and we made sure that the participant is able to complete the search goal as long as he/she examines all the provided search results. Each time the participant completed a search session, he/she was required to label a satisfaction score to reflect his/her search experience. Then they would be guided to continue to the next search task.

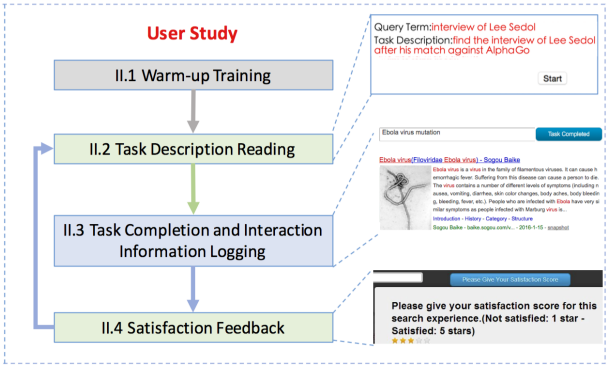

Figure 1: Data Collection Procedure

No query reformulation are allowed and the number of search results are fixed to 10 for the consistency of result sets across users. Meanwhile, users' satisfaction feedback is collected at SERP-level rather than at session-level because most offline metrics are designed to measure the quality of only one search result page.

2.2 Task Organization of Dataset #1

Dataset #1 is the data used in [4] and the search results are all organic results. The 26 search tasks were selected from NTCIR IMine task [5]. All the queries were collected from a commercial search engine and were neither long-tailed nor hot ones. Different from the IMine task, both queries and detailed task explanations were provided to the participants to avoid ambiguity.

For each search task, the query and results are fixed to ensure the data consistency. The search results were collected from a popular commercial search engine and only top 10 organic results

Table 2: Examples of Search Tasks and Manipulated Off-target Queries to Retrieve Verticals

Vertical Presentation Style	Original Query	Off-target Query
Textual	poems describing spring rain	poems describing rain
	ancient Greek architectural style	ancient Greek
Encyclopedia	covering the sky (novel)	covering the sky (game)
	the 9^{th} zone (movie)	the 9^{th} zone (novel)
Image	nike basketball shoes	nike football shoes
	pictures of wine cabinet	pictures of cupboard
Download	iTunes download	iTools download
	Renren desktop app download	Weibo desktop app download
Encyclopedia	ebola virus mutation	news of ebola virus
	Chinese city competitiveness	Chinese enterprise competitiveness

are retained. Vertical results and advertisements were not included in this dataset.

Due to the research purpose of study users' variability in satisfaction perception, the SERPs were manipulated in Dataset #1. Three different SERPs are designed for each search task based on the result relevance annotations, namely the "ordered-page", "reversed-page" and "random-page". For each query, the results on three SERPs are the same but in different ranking orders. On the "ordered-page" and "reversed-page", the results were ranked in the order/reverse order of relevance, respectively. On the "random-page", the results were ranked in a random order. In this case, there are 3 different rankings per query.

2.3 Task Organization of Dataset #2

Dataset #2 is mainly the data used in [1] and the search results contain a number of vertical results. The 30 queries in Dataset #2 are sampled from the search logs from a major commercial search engine and are neither long-tailed nor popular ones. With these queries, search tasks were organized and on/off-topic verticals as well as the non-vertical results were crawled from the commercial search engine.

Due to the research purpose of studying the effect of vertical results on search satisfaction, the SERPs for each search task vary in three aspects:

- Quality: The vertical results in Dataset #2 contain on-topic ones and off-topic ones. The on-topic verticals are crawled from the commercial search engine with the corresponding query. For the off-topic verticals, we use a subset of the terms from the original query or add a few new items to generate an off-target query. With the off-target query, off-topic verticals can be crawled from the commercial search engine. Because the new query just overlapped a subset of the original one, the vertical results obtained are usually irrelevant to the original query but appear to be quite similar to the on-topic verticals. Table 1 shows some examples of the search tasks and corresponding queries used to crawl on/off-topic verticals in our experiment.
- Position: The vertical results were randomly placed at position 1, 3 and 5 of the search result lists.
- Presentation styles: The 30 search tasks contain 5 types of verticals, namely, textual vertical, image vertical, news vertical, download vertical and encyclopedia vertical results. For each presentation style, we have 6 search tasks.

In this way, there were six (2 quality types $\times$ 3 position ranks) different SERPs for each seach task. Each SERP was composed of one vertical result and nine non-vertical results. The non-vertical results were aslo crawled from the same commercial search engine, and were kept the original orders unchanged.

Part of the tasks in Dataset #2 are the same as that in Dataset #1 described in Section 3. As a result, there are 9 different SERPs (6 heterogeneous pages in Dataset #2 + 3 homogeneous pages in Dataset #1) for these tasks.

Moreover, for some of the tasks in Dataset #2, there is an extra SERP which contains the top 10 results crawled from the commercial search engine without any manipulation. There are 7.4 vertical results on these tasks on average. In this way, some tasks may have seven differentSERPs (6 single-vertical SERPs + 1 multi-vertical SERP) or ten different SERPs (6 single-vertical pages + 3 homogeneous pages + 1 multi-vertical page).

2.4 Participants

We invited 40/58 participants for the data collection process of Dataset #1/#2, respectively. Each of the participants was required to complete roughly 30 search tasks. During the experiment procedure, we adopted a Graeco-Latin square design and randomized sequence order to ensure that each task condition had the same opportunity to be shown to users.

2.5 Relevance Annotation

Three professional assessors from a commercial search engine company were invited to label the relevance scores for all query-result pairs. The KAPPA coefficient of the their annotation is 0.70, which can be characterized as a substantial agreement according to Cohen [3].

3 CONCLUSIONS

The datasets used in the SIGIR17 paper "Meta-evaluation of Online and Offline Web Search Evaluation Metrics" are composed of more than 2400 search sessions collected from 98 users. With the rich information provided by the datasets, the authors could meta-evaluate online/offline evaluation metrics and study how they infer

actual user satisfaction. This paper provides detailed description of the data generation process of the datasets used in [2].

REFERENCES

[1] Ye Chen, Yiqun Liu, Ke Zhou, Meng Wang, Min Zhang, and Shaoping Ma. 2015. Does Vertical Bring more Satisfaction?: Predicting Search Satisfaction in a Heterogeneous Environment. In *CIKM'15*. ACM, 1581–1590.

[2] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of Online and OfilineWeb Search Evaluation Metrics. In *SIGIR'17*. ACM.

[3] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.

[4] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR'15*. ACM, 493–502.

[5] Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto Kato, Hiroaki Ohshima, and Ke Zhou. 2014. Overview of the NTCIR-11 IMine task. In *NTCIR*, Vol. 14.