Overview of the NTCIR-14 OpenLiveQ-2 Task

Makoto P. Kato[1], Akiomi Nishida[2], Tomohiro Manabe[2], Sumio Fujita[2], and Takehiro Yamamoto[1]

[1] Kyoto University mpkato@acm.org, tyamamot@dl.kuis.kyoto-u.ac.jp
[2] Yahoo Japan Corporation {tomanabe,anishida,sufujita}@yahoo-corp.jp

Abstract. This is an overview of the NTCIR-14 OpenLiveQ-2 task. This task aims to provide an open live test environment of Yahoo Japan Corporation's community question-answering service (*Yahoo! Chiebukuro*) for question retrieval systems. The task was simply defined as follows: given a query and a set of questions with their answers, return a ranked list of questions. Submitted runs were evaluated both offline and online. In the online evaluation, we employed *pairwise preference multileaving*, a multileaving method that showed high efficiency over the other methods in a recent study. We describe the details of the task, data, and evaluation methods, and then report official results at NTCIR-14 OpenLiveQ-2.

Keywords: online evaluation · interleaving · community question answering

1 Introduction

Community Question Answering (cQA) services are Internet services in which users can ask a question and obtain answers from other users. Users can obtain relevant information to their search intents not only by asking questions in cQA, but also by searching for questions that are similar to their intents. Finding answers to questions similar to a search intent is an important information seeking strategy especially when the search intent is very specific or complicated. While a lot of work has addressed the question retrieval problem [7, 1, 8], there are still several important problems to be tackled:

Ambiguous/underspecified queries Most of the existing work mainly focused on specific queries. However, many queries used in cQA services are as short as Web search queries, and, accordingly, ambiguous/underspecified. Thus, question retrieval results also need diversification so that users with different intents can be satisfied.

Diverse relevance criteria The notion of relevance used in traditional evaluation frameworks is usually *topical relevance*, which can be measured by the degree of match between topics implied by a query and ones written in a document. Whereas, real question searchers have a wide range of relevance criteria such as freshness, concreteness, trustworthiness, and conciseness. Thus, traditional relevance assessment may not be able to measure real performance of question retrieval systems.

In order to address these problems, we have organized a task called *Open Live Test for Question Retrieval* (*OpenLiveQ*) since 2016, which provides an open live test environment of Yahoo! Chiebukuro[3] (a Japanese version of Yahoo! Answers) for question retrieval systems. Participants can submit ranked lists of questions for a particular set of queries, and receive evaluation results based on real user feedback. Involving real users in evaluation can solve problems mentioned above: we can consider the diversity of search intents and relevance criteria by utilizing real queries and feedback from users who are engaged in real search tasks.

The NTCIR-14 OpenLiveQ-2 task is the second round of OpenLiveQ. The most of the settings in OpenLiveQ-2 are the same as those in the first round of OpenLiveQ (OpenLiveQ-1) [3]. We used the same task definition and the same query set for both training and testing, while we updated questions to be retrieved and clickthrough data in OpenLiveQ-2, and employed a new evaluation methodology for evaluating a large number of runs. In OpenLiveQ-1, only selected runs were evaluated in the online evaluation, since a prohibitively large amount of impressions were expected to statistically distinguish all the submitted runs. OpenLiveQ-2 tried to address this problem by proposing two-phase online evaluation [2]: the first phase identifies top-k runs and the second phase finds statistically significant differences only among top-k runs.

The remainder of the paper is organized as follows. Section 2 describes the OpenLiveQ-2 task in details. Section 3 introduces the data distributed to OpenLiveQ-2 participants. Section 4 explains a new evaluation methodology applied to the OpenLiveQ-2 task.

2 Task

The task of the OpenLiveQ-2 task is simply defined as follows: given a query and a set of questions with their answers, return a ranked list of questions. Our task consists of three phases:

1. **Offline Training Phase** Participants were given *training data* including a list of queries, a set of questions for each query, and clickthrough data (see Section 3 for details). They could develop and tune their question retrieval systems based on the training data.
2. **Offline Test Phase** Participants were given only a list of queries and a set of questions for each query. They were required to submit a ranked list of questions for each query by a deadline. We evaluated submitted results by using evaluation metrics for ad-hoc retrieval with relevance judgment data that we developed in OpenLiveQ-1. Unlike OpenLiveQ-1, the results of the offline evaluation were only used for excluding poor ranking results that can drastically degrade the user satisfaction during the online evaluation. Meanwhile, we did not exclude any submitted runs in OpenLiveQ-2 since no run underperformed baseline runs to a large extent.

[3] http://chiebukuro.yahoo.co.jp/

3. **Online Test Phase** All the submitted runs were evaluated in a production environment of Yahoo Japan Corporation. A *multileaved comparison* method [4] was used in the online evaluation. As briefly mentioned in Section 1, OpenLiveQ-2 employed the two-phase online evaluation for evaluating a large number of runs efficiently.

As the open live test is conducted on a Japanese service, the language scope is limited to Japanese. Meanwhile, we supported participants by providing a tool for feature extraction so that Japanese NLP is not required for participation.

3 Data

This section explains the data used in the OpenLiveQ-2 task. Note that we omit explanation about the query set since we used the same query set as that in NTCIR-13 OpenLiveQ-1 [3].

3.1 Questions

Questions were prepared in the same way as that in OpenLiveQ-1. We input each query to the current Yahoo! Chiebukuro search system as of Apr 10, 2018, recorded the top 1,000 questions, and used them as questions to be ranked. Information about all the questions as of Apr 10, 2018 was distributed to the OpenLiveQ participants, and includes

- Query ID (a query by which the question was retrieved),
- Rank of the question in a Yahoo! Chiebukuro search result for the query of Query ID,
- Question ID,
- Title of the question,
- Snippet of the question in a search result,
- Status of the question (accepting answers, accepting votes, or solved),
- Last update time of the question,
- Number of answers for the question,
- Page view of the question,
- Category of the question,
- Body of the question, and
- Body of the best answer for the question.

The total number of questions is 1,971,816. As was mentioned earlier, participants were required to submit a ranked list of those questions for each test query.

3.2 Clickthrough Data

Clickthrough data were collected in the same way as that in OpenLiveQ-1, and available for some of the questions. Based on the clickthrough data, one can

estimate the click probability of the questions, and understand what kinds of users click on a certain question. The clickthrough data were collected from Jan 10, 2018 to Apr 9, 2018.

The clickthrough data include

- Query ID (a query by which the question was retrieved),
- Question ID,
- Most frequent rank of the question in a Yahoo! Chiebukuro search result for the query of Query ID,
- Clickthrough rate,
- Fraction of male users among those who clicked on the question,
- Fraction of female users among those clicked on the question,
- Fraction of users under 10 years old among those who clicked on the question,
- Fraction of users in their 10s among those who clicked on the question,
- Fraction of users in their 20s among those who clicked on the question,
- Fraction of users in their 30s among those who clicked on the question,
- Fraction of users in their 40s among those who clicked on the question,
- Fraction of users in their 50s among those who clicked on the question, and
- Fraction of users over 60 years old among those who clicked on the question.

The clickthrough data contain click statistics of a question identified by Question ID when a query identified by Query ID was submitted. The rank of the question can change even for the same query. This is why the third value indicates the most frequent rank of the question. The number of query-question pairs in the clickthrough data is 436,890.

4 Evaluation

This section describes submissions from NTCIR-14 OpenLiveQ-2 participants, and then introduces the offline evaluation, in which runs were evaluated with relevance judgment data, and online evaluation, in which runs were evaluated with real users by means of multileaving.

4.1 Submissions

The NTCIR-14 OpenLiveQ-2 task attracted five research teams including an organizer team. The total number of submitted runs during the offline test phase was 65, of which 4 runs were duplicates of the other submissions. Thus, the total number of unique runs was 61.

4.2 Offline Evaluation

The offline evaluation was conducted in a similar way to traditional ad-hoc retrieval tasks, in which results are evaluated by relevance judgment results and evaluation metrics such as nDCG (normalized discounted cumulative gain), ERR (expected reciprocal rank), and Q-measure. During the offline test period,

participants could submit their results once per day through our Web site[4], and obtain evaluation results right after the submission.

While test questions used in OpenLiveQ-2 were not exactly the same as those in OpenLiveQ-1, we reused relevance judgment data in OpenLiveQ-2. The number of judged test questions was 43,205, *i.e.* 4.38% of all the test questions in OpenLiveQ-2. This fraction is comparable to that in OpenLiveQ-1, in which 4.54% of questions were judged. We used *condensed list* approach [5] to deal with incomplete relevance judgment data, *i.e.* we filtered out questions without relevance judgments from ranked lists of submitted runs.

Only a score of Q-measure for each submitted run was displayed at our website. This is primarily because our recent study showed high correlation between the Q-measure scores and online evaluation results [2].

4.3 Online Evaluation

The NTCIR-13 OpenLiveQ-1 task attracted seven research teams and received 85 submissions in total. Even though the multileaved comparison can evaluate multiple rankings simultaneously, a large amount of search results impressions are required for a large number of rankers according to simulation-based experiments [4]. Thus, we opted to select a subset of submitted rankers by means of offline evaluation, and conducted multileaved comparison for only ten selected rankers — it turned out to be a problematic experimental design.

Lessons from NTCIR-13 OpenLiveQ task are summarized as follows [2]: (1) The offline evaluation result in terms of Q-measure [6] showed high correlation to the online evaluation result. However, there were some rankers for which the offline and online evaluation strongly disagreed. This result implies a potential problem of our strategy: we might not evaluate rankers better than those selected for the online evaluation. This is a serious problem not only for an evaluation campaign but also for improvement of Web services. A straightforward solution to this problem is to evaluate all the rankers online. (2) A large number of users' clicks were necessary to find statistically significant differences for all the ranker pairs. As we cannot easily increase the number of search result impressions for multileaved comparison, a straightforward solution to this problem is to evaluate less rankers online.

These lessons motivated us to devise a new experimental design for large-scale multileaved comparison. Our proposed methodology in OpenLiveQ-2, two-phase online evaluation, is to evaluate all the rankers online for identifying top-k rankers, and intensively compare the top-k rankers so that they can get more chances to be statistically distinguished. We tested several top-k identification methods for multileaved comparison based on simulation-based experiments in our recent study [2]. The results demonstrated that even a simple method, Copeland counting algorithm, could achieve high recall in the top-k identification problem. Thus, OpenLiveQ-2 employed the two-phase online evaluation

[4] http://www.openliveq.net/

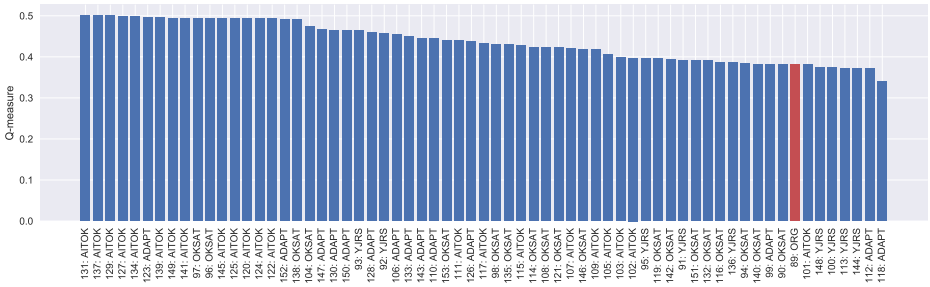

Fig. 1. Offline evaluation: Q-measure.

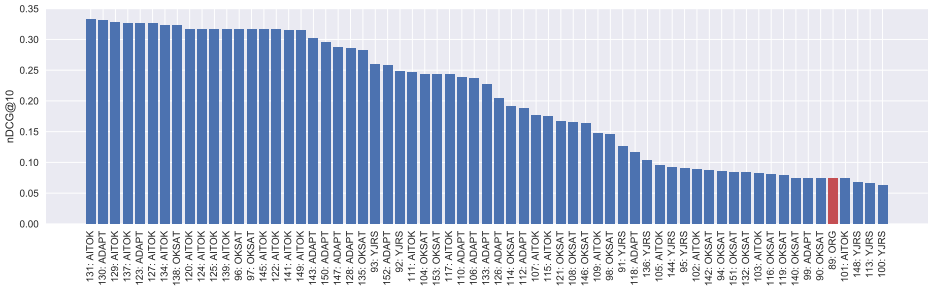

Fig. 2. Offline evaluation: nDCG@10.

for evaluating all the submitted runs, with a recently proposed multileaving algorithm, pairwise preference multileaving [4].

In our online evaluation, we evaluated only 61 unique runs out of 65 after excluding duplicate runs. The first phase of the two-phase online evaluation was carried out from Sep 28, 2018 to Nov 11, 2018. The total number of impressions used was 164,478 at the first phase. After identifying top-k runs based on the results of the first phase ($k = 30$ in OpenLiveQ-2), we evaluated only those top runs from Nov 23, 2018 to Jan 6, 2019. The total number of impressions used was 148,976 at the second phase.

5 Evaluation Results

Figures 1, 2, and 3 show results of the offline evaluation in terms of Q-measure, nDCG@10, and ERR@10. The baseline run (89: ORG) are indicated in red and was produced exactly the same ranked list as that used in the production.

Figures 4 and 5 show cumulated credits in the online evaluation at the first and second phase, respectively. Note that *the official result of NTCIR-14 OpenLiveQ-2 is that at the second phase, and online evaluation result at the first phase is only considered as unofficial due to lack of statistical power.*

Table 1 shows results of Tukey's HSD tests with $\alpha = 0.05$ at the second phase. $*$ indicates statistical significance of a run pair.

Table 1. Results of Tukey's HSD tests with $\alpha = 0.05$. $*$ indicates statistical significance of a run pair.

	93	95	100	102	104	105	106	107	109	111	112	113	115	117	118	122	124	126	127	129	131	134	135	137	138	139	141	145	147
92				*		*		*	*		*		*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*
93	−			*		*								*		*	*		*	*	*	*	*	*	*	*	*	*	*
95	−	−		*												*	*		*	*	*	*		*	*		*	*	*
100	−	−	−	*		*		*	*		*		*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*
102	−	−	−	−			*			*		*			*														
104	−	−	−	−	−																				*				
105	−	−	−	−	−	−				*		*			*														
106	−	−	−	−	−	−	−									*	*		*	*	*	*		*	*	*	*	*	*
107	−	−	−	−	−	−	−	−				*																	
109	−	−	−	−	−	−	−	−	−			*																	
111	−	−	−	−	−	−	−	−	−	−				*		*	*		*	*	*	*	*	*	*	*	*	*	*
112	−	−	−	−	−	−	−	−	−	−	−	*																	
113	−	−	−	−	−	−	−	−	−	−	−	−	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*
115	−	−	−	−	−	−	−	−	−	−	−	−	−																
117	−	−	−	−	−	−	−	−	−	−	−	−	−	−															
118	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	*	*		*	*	*	*	*	*	*	*	*	*	*
122	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−													
124	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−												
126	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−											
127	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−										
129	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−									
131	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−								
134	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−							
135	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−						
137	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−					
138	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−				
139	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−			
141	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		
145	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	

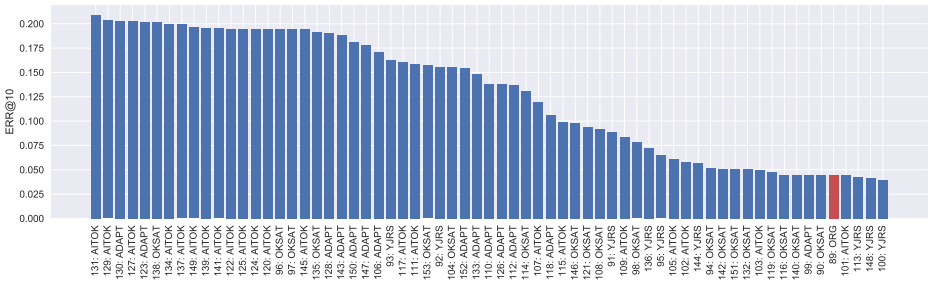

Fig. 3. Offline evaluation: ERR@10.

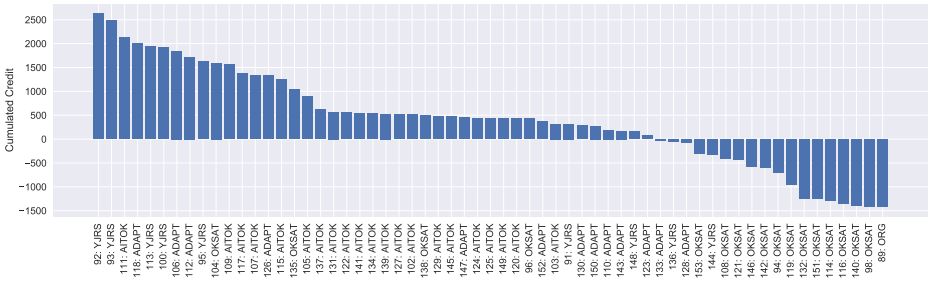

Fig. 4. Online evaluation at the first phase.

References

1. Cao, X., Cong, G., Cui, B., Jensen, C.S.: A generalized framework of exploring category information for question retrieval in community question answer archives. In: WWW. pp. 201–210 (2010)
2. Kato, M.P., Manabe, T., Fujita, S., Nishida, A., Yamamoto, T.: Challenges of multileaved comparison in practice: Lessons from ntcir-13 openliveq task. In: CIKM. pp. 1515–1518 (2018)
3. Kato, M.P., Yamamoto, T., Manabe, T., Nishida, A., Fujita, S.: Overview of the ntcir-13 openliveq task. In: NTCIR-13 Conference (2017)
4. Oosterhuis, H., de Rijke, M.: Sensitive and scalable online evaluation with theoretical guarantees. In: CIKM. pp. 77–86 (2017)
5. Sakai, T.: Alternatives to bpref. In: SIGIR. pp. 71–78 (2007)
6. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: SIGIR. pp. 1043–1052 (2011)
7. Wang, K., Ming, Z., Chua, T.S.: A syntactic tree matching approach to finding similar questions in community-based qa services. In: SIGIR. pp. 187–194 (2009)
8. Zhou, G., Liu, Y., Liu, F., Zeng, D., Zhao, J.: Improving question retrieval in community question answering using world knowledge. In: IJCAI. pp. 2239–2245 (2013)

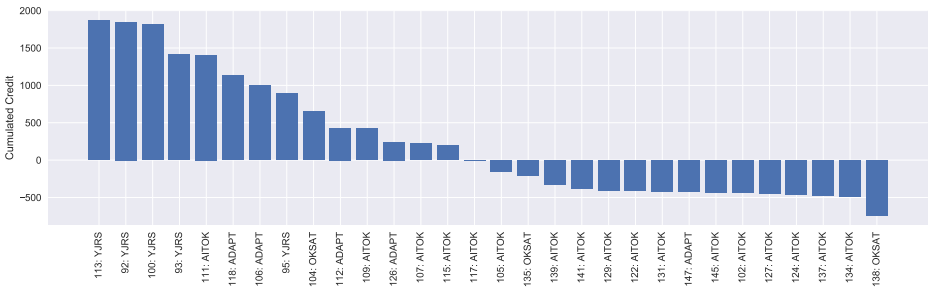

Fig. 5. Online evaluation at the second phase.