Overview of the NTCIR-16 Data Search 2 Task

Makoto P. Kato
University of Tsukuba
mpkato@acm.org

Hiroaki Ohshima
University of Hyogo
ohshima@ai.u-hyogo.ac.jp

Ying-Hsang Liu
University of Southern Denmark
yingliu@sdu.dk

Hsin-Liang Chen
Missouri University of Science and Technology
chenhs@mst.edu

ABSTRACT

NTCIR-16 Data Search 2 is the second round of the Data Search task at NTCIR. The first round of Data Search (NTCIR-15 Data Search) focused on the retrieval of a statistical data collection. This round also addressed the problem of ad-hoc data retrieval (IR subtask) and planned the other subtasks including question answering (QA) subtask and user interface (UI) subtask. This paper introduces the task definition, test collection, and evaluation methodology of the subtasks of NTCIR-16 Data Search 2. The IR subtask attracted seven research groups, from which we received 25 English runs and 23 Japanese runs. The evaluation results of these runs are presented and discussed in this paper.

1 INTRODUCTION

The open data movement is now being accelerated by the expectations for open science and citizen science. It is said that researchers all over the world could collaborate on world-wide problems and citizens also could participate in research activities if various kinds of data were publicly available. The government of each country strongly encourages the open data movement and has launched open-data government initiatives such as Data.gov[1] in the United States, Data.gov.uk in the United Kingdom, and e-Stat[2] in Japan. Besides the governmental portals, there are also thousands of data repositories on the Web [3].

The growth of the open data movement has naturally motivated researchers and industries to develop search engines for the open data scattered on the Web. Google launched Google Dataset Search as public beta in September 2018 [4], and some researchers have started to discuss potential research topics of data search [1]. Although there have been several attempts for understanding and developing data search, neither a benchmark nor an evaluation campaign on data search has not been proposed yet.

Following rapidly increasing demands and interests in data search, we proposed a pilot task on data search, *Data Search*, at NTCIR-15 [2]. The first round of Data Search focused on the retrieval of a statistical data collection published by the Japanese government (e-Stat), and one published by the US government (data.gov). NTCIR-15 Data Search attracted six research groups and received 17 submissions for the Japanese subtask, and 37 submissions for the English subtask.

Lessons learned at the NTCIR-15 Data Search task are summarized as follows:

L1. The inter-rater agreement of relevance judgments was not high enough. This was partially because we did not enforce assessors to investigate statistical data in depth. More careful assessments would be required to produce reliable relevance judgments.

L2. We took the maximum effectiveness measure for each query and produced an *oracle* run. There was a large gap between the oracle and participants' runs. It may suggest that there is much room for improvement in this task.

L3. The top three runs show very different performances for each query. No single run could achieve satisfactory results for all the queries. More insights into this task would be required to advance data search systems.

In addition to the lessons from NTCIR-15, there still remain several challenges at the previous round. Even though the data search system can return relevant datasets in response to a given query, users have (**C1**) to identify relevant datasets in the search engine result page, and (**C2**) to locate relevant parts within a dataset, both of which requires a considerable amount of users' efforts. Therefore, the ad-hoc retrieval task at the first round alone might not be sufficient for building data search systems.

Building on the lessons and limitations discussed at NTCIR-15 Data Search, we proposed the second round of Data Search at NTCIR-16, which consists of three subtasks, namely, ad-hoc retrieval, question answering, and search interface subtasks. The ad-hoc retrieval subtask is a standard ad-hoc retrieval task for statistical data and inherits the task design at the first round, aiming to address the lessons **L2** and **L3**. The question answering subtask is similar to reading comprehension tasks, in which participants are required to extract answers from statistical data for a given question. This subtask was designed to address **L1** and **C2**: the extracted answers in this subtask can be useful resources for judging the relevance of datasets in the ad-hoc retrieval subtask, and this question answering system would save users' efforts identifying relevant parts within a dataset. The search interface subtask is an exploratory subtask in which participants are required to develop a search system with an effective search interface for data search tasks. We do not have an explicit evaluation for this task, and expect participants to present their demonstrations at the NTCIR conference. The challenge **C1** would be addressed by this subtask.

In the remainder of this paper, we introduce the task design, resources, evaluation methodology, and evaluation results of each subtask.

[1] https://www.data.gov/
[2] https://www.e-stat.go.jp/

Table 1: Statistics of the available resources at the NTCIR-16 Data Search 2 IR subtask.

Subtask	Resource	#
Japanese		
	Datasets	1,338,402
	Data files	1,338,402
	NTCIR-15 Queries	192
	NTCIR-15 qrels	7,754
English		
	Datasets	46,615
	Data files	92,930
	NTCIR-15 queries	192
	NTCIR-15 qrels	8,248

2 IR SUBTASK

This section introduces information retrieval (IR) subtask of NTCIR-16 Data Search 2.

2.1 Task

The task of the IR subtask is the same as that in the NTCIR-15 Data Search task and is defined as follows: Given a query for data search, a system is expected to return a ranked list of *datasets*.

The definition of dataset is also the same as that in the previous round: a pair of metadata and a set of data files. Data collections used in the Data Search tasks are (1) data.gov data collection, where the same metadata is used for multiple data files (i.e., there is one-to-many relationship between metadata and data files), and (2) e-Stat data collection, where metadata is given for individual data files. Figure 1 shows an example of a dataset in the data.gov data collection, which consists of multiple data files, i.e., an Excel file and a CSV file. Metadata constitutes pairs of an attribute and its value. For example, attributes “Metadata Created Date” and “Metadata Updated Date” are included in the example in Figure 1. Data files are Excel (i.e., xls and xlsx), CSV, PDF, XML, JSON, RDF, and text files for the data.gov data collection, and Excel, CSV, and PDF files for the e-stat data collection.

In NTCIR-16 Data Search 2, each team was allowed to submit up to 10 runs. Runs should be generated automatically.

2.2 Resources

In NTCIR-16 Data Search 2, the participants were allowed to use topics and relevance judgments developed in the NTCIR-15 Data Search task. The statistics of the available resources are shown in Table 1.

Since the most of the resources were the same as those in NTCIR-15, we only explain new resources developed in the current round.

2.2.1 Topics. In the NTCIR-15 Data Search task, we developed topics based on questions posted in a community question answering service. As some users post an inquiry about datasets, we gathered such questions and manually extracted realistic information needs for data search. Since there did not remain “data search” questions for developing new topics, we looked for another resource at this round. The resource for the topic development at this round was the webpage referring to datasets. We first parsed commoncrawl[3] webpages crawled from August 2018 to April 2019 (approximately 25 billion web pages), and identified 47,242 URLs including “data.gov” and 137,388 URLs including “stat.go.jp”. We read some of the webpages containing those URLs and manually created information needs assuming that the user has a question that can be answered by the webpage content. For example, if a webpage describes the most popular white names by referring to a dataset on popular names in the US, we created an information need that requires the page content as an answer, for example, “What are the most common white names in 2012?”. Some examples of the information needs are shown in Table 2.

```
{
    "id": "0063664a-d0d7-4ce2-9462-0463a89fc274",
    "url": "https://catalog.data.gov/dataset/0063664a-
         d0d7-4ce2-9462-0463a89fc274",
    "attribution": "CRED REA Fish Team Stationary Point
         Count Surveys at Sarigan, Marianas Archipelago,
         2005 (https://catalog.data.gov/dataset/0063664a-
         d0d7-4ce2-9462-0463a89fc274) is licensed under U
         .S. Government Work (http://www.usa.gov/
         publicdomain/label/1.0/)"
    "title": "CRED REA Fish Team Stationary Point Count
         Surveys at Sarigan, Marianas Archipelago, 2005",
    "description": "Stationary Point Counts at 4 stations
          at each survey site were surveyed as part of
         Rapid Ecological Assessments (REA) conducted at
         3 sites around Sarigan in the Marianas
         Archipelago (MA) during 3 September - 1 October
         2005 in the NOAA Oscar Elton Sette (OES 0511)
         Reef Assessment and Monitoring Program (RAMP)
         Cruise. Raw survey data included species level
         abundance estimates.",
    "data": [
        {
            "data_format": "excel",
            "data_organization": "National Oceanic and
                 Atmospheric Administration, Department
                 of Commerce",
            "data_url": "https://data.nodc.noaa.gov/coris
                 /data/NOAA/nmfs/pifsc/cred/REAFish/
                 CNMI_2005/CRED_REA_FISH_SAIPAN_2005.xls"
                 ,
            "data_filename": "CRED_REA_FISH_SAIPAN_2005.
                 xls"
        },
        {
            "data_format": "csv",
            ...
        }
    ],
    "data_fields": {
        "Resource Type": "Dataset",
        "Metadata Date": "June 20, 2018",
        "Metadata Created Date": "February 7, 2018",
        "Metadata Updated Date": "February 27, 2019",
        ...
        "metadata_sources": [
            "https://catalog.data.gov/harvest/object/
                 fc5a39b7-4c9f-49b8-af95-2812d9b3264c"
        ]
    }
}
```

Figure 1: Example of metadata of an English dataset.

[3] https://commoncrawl.org/

Table 2: Examples of the topics and queries.

Topic ID	Information need	Query
DS2-E-0001	What is the largest causes of death in United State in 1999-2016?	causes of death us 1999-2016
DS2-E-0002	Where is the most active tectonic faulting region near Mount Rainier?	most active tectonic fault mount rainier
DS2-E-0003	Where is the nearest park to my office?	office nearest park
DS2-E-0004	How much is the tuition fee at a private elementary school?	tuition fee private elementary school
DS2-E-0005	Are there hospital differences across the US states?	us states hospital differences

The number of topics was determined with the guidance of topics size design methodology proposed by Sakai [5]. We first computed the residual variance of nDCG@10 based on the NTCIR-15 Data Search results: $\sigma^2 = 0.00281$ for English runs and $\sigma^2 = 0.00540$ for Japanese runs. These values were used in samplesizeANOVA2.xlsx[4], where minD = 0.05 (the minimum detectable difference), $m = 10$ (the number of systems), $\alpha = 0.05$, and $\beta = 0.20$. As a result, we obtained $n = 36$ for English runs and $n = 68$ for Japanese runs. This indicates that n topics were enough for achieving 80% power for finding statistical differences of 0.05 or higher in terms of nDCG@10 for m systems. The number of the developed topics was larger than those expected numbers: there are 58 topics for English and 72 topics for Japanese at the NTCIR-16 Data Search 2 IR subtask.

Queries were obtained from information needs in the same way as that in the NTCIR-15 Data Search task. We asked ten crowdsourcing workers to generate queries based on presented information needs assuming that they were searching for the answer to the question in the information need. Gathering users' generated queries, we selected the most representative query for each topic (See the overview paper for more detail [2]). Some examples of the queries are shown in Table 2, together with their information needs.

2.2.2 Collection. In NTCIR-16 Data Search 2, we used exactly the same dataset collections as those used in NTCIR-15 Data Search, i.e., data.gov and e-Stat data collections. More precisely, `data_search_e_collection.jsonl.bz2` (metadata) and `data_search_e_data.tar.bz2` (data files) were used as the English dataset collection, while `data_search_j_collection.jsonl.bz2` (metadata) and `data_search_j_data.tar.bz2` (data files) were used as the Japanese dataset collection. These files are available at https://ntcir.datasearch.jp/data/. Figure 1 shows an example of a dataset in the data.gov data collection.

2.2.3 Baseline Systems. We implemented several standard baseline systems such as BM25, LM, and BM25+RM3 by using Anserini [6]. All the baseline systems regarded each dataset as a unstructured document containing texts in "title" and "description" fields of the metadata. They were available to the participants[5].

2.2.4 Relevance Judgments. Relevance judgments were conducted in exactly the same way as those in the NTCIR-16 Data Search 2 task. We pooled the top 10 results from each run, and evaluated the relevance grade by crowd-sourcing services. Amazon Mechanical Turk[6] was used for English runs, while Lancers[7] was used for Japanese runs. Each topic-dataset pair was evaluated at a three-point scale (0: irrelevant, 1: partially relevant, and 2: highly relevant).

We assigned five workers to each topic-dataset pair for Japanese runs, while three assessors were assigned for English runs with an option "Require that Workers be Masters to do your tasks". This setting is identical to that in the previous round [2]. Krippendorff's α is 0.444 for English and 0.474 for Japanese. These are fairly consistent with the inter-rator agreements in the NTCIR-15 Data Search task: 0.438 for English and 0.478 for Japanese.

2.3 Evaluation Results

The IR subtask attracted seven research groups, from which we received 25 English runs and 23 Japanese runs. All the submitted runs are listed in Tables 3 and 4. Each run was named "[GROUP_ID]-[LANGUAGE]-[PRIORITY]" where "[GROUP_ID]" is a group ID, "[LANGUAGE]" is either "E" (English) or "J" (Japanese), and "[PRIORITY]" is an integer between 1 and 10, indicating which runs should be prioritized in the pooling for relevance assessments. Each run file was required to include a system description, which is also shown in the table. The last column of each run table indicates features of the system declared by each team. The value of this column is defined as [DATA],[NEURAL],[ENTITY],[NUMBER], where [DATA] indicates whether the data files are used (Y (Yes) or N (No)), [NEURAL] indicates whether neural language models (e.g., BERT, RoBERT, GPT, and T5) are used, [ENTITY] indicates whether entities are treated differently from the other tokens, and [NUMBER] indicates whether numbers are treated differently from the other tokens.

Tables 5 and 6 show the evaluation results of English and Japanese IR subtask runs, respectively. Runs are sorted by nDCG@10, which is our primary evaluation metric.

3 QA SUBTASK

This task can be considered as an extension of the ad-hoc retrieval subtask: given a question about statistical data, a system is expected to extract an answer to the question. As we mentioned earlier, some of the topics used in the ad-hoc retrieval subtask will be used as questions.

4 UI SUBTASK

The user interface subtask is an exploratory attempt at the second round. This subtask requires participants to develop a search system with an effective search interface for data search tasks.

[4] http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx
[5] Available at https://github.com/mpkato/ntcir-datasearch
[6] https://www.mturk.com/
[7] https://www.lancers.jp/

Table 3: Runs submitted to the NTCIR-16 Data Search IR subtask (English).

Run name	Description	Run type
KSU-E-1	Category+Table Clipping+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-3	Category+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-5	Table Clipping+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-7	Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-9	Category+Table Header+BM25	Y,N,Y,N
NYUCIN-E-1	BM25 and BERT	Y,Y,N,N
ORGE-E-1	bm25prf+bm25	N,N,N,N
ORGE-E-2	bm25	N,N,N,N
ORGE-E-3	bm25.accurate	N,N,N,N
ORGE-E-4	sdm+qld	N,N,N,N
ORGE-E-5	rm3+bm25	N,N,N,N
ORGE-E-6	qld	N,N,N,N
ORGE-E-7	sdm+bm25	N,N,N,N
ORGE-E-8	rm3+qld	N,N,N,N
OUHCIR-E-1	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N,Y,N,N
OUHCIR-E-2	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N,Y,N,N
OUHCIR-E-3	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-4	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-5	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-6	BM25 and TFIDF WEIGHT ADJUSTED	NNNN
OUHCIR-E-7	DOC2VEC	N,N,N,N
OUHCIR-E-8	DOC2VEC	NNNN
STIS-E-1	prop+bert_score+bm25	Y,Y,N,N
STIS-E-2	prop+bert_score	Y,Y,N,N
wut21-E-1	LM Jelinek Mercer	Y,N,N,N

5 CONCLUSIONS

This paper introduces the task definition, test collection, and evaluation methodology of the subtasks of NTCIR-16 Data Search 2. The IR subtask attracted seven research groups, from which we received 25 English runs and 23 Japanese runs.

ACKNOWLEDGMENTS

We would like to thank the Data Search participants for their contributions to the NTCIR-16 Data Search 2 task. We also thank the NTCIR office to support our activities.

REFERENCES

[1] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis Daniel Ibáñez-Gonzalez, Emilia Kacprzak, and Paul T. Groth. 2019. Dataset search: a survey. *CoRR* abs/1901.00735 (2019). http://arxiv.org/abs/1901.00735

[2] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2020. Overview of the NTCIR-15 Data Search Task. In *NTCIR-15*.

[3] D-Lib Magazine. 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* 23, 3/4 (2017).

[4] Natasha Noy, Matthew Burgess, and Dan Brickley. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *WebConf.* 1365–1375.

[5] Tetsuya Sakai. 2016. Topic set size design. *Information Retrieval Journal* 19, 3 (2016), 256–283.

[6] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *Journal of Data and Information Quality* 10, 4, Article 16 (Oct. 2018), 20 pages. https://doi.org/10.1145/3239571

Table 4: Runs submitted to the NTCIR-16 Data Search IR subtask (Japanese).

Run name	Description	Run type
KSU-E-1	Category+Table Clipping+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-3	Category+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-5	Table Clipping+Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-7	Table Header+BERT+MLP	Y,Y,Y,N
KSU-E-9	Category+Table Header+BM25	Y,N,Y,N
NYUCIN-E-1	BM25 and BERT	Y,Y,N,N
ORGE-E-1	bm25prf+bm25	N,N,N,N
ORGE-E-2	bm25	N,N,N,N
ORGE-E-3	bm25.accurate	N,N,N,N
ORGE-E-4	sdm+qld	N,N,N,N
ORGE-E-5	rm3+bm25	N,N,N,N
ORGE-E-6	qld	N,N,N,N
ORGE-E-7	sdm+bm25	N,N,N,N
ORGE-E-8	rm3+qld	N,N,N,N
OUHCIR-E-1	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N,Y,N,N
OUHCIR-E-2	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N,Y,N,N
OUHCIR-E-3	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-4	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-5	BM25 and TFIDF WEIGHT ADJUSTED	N,N,N,N
OUHCIR-E-6	BM25 and TFIDF WEIGHT ADJUSTED	NNNN
OUHCIR-E-7	DOC2VEC	N,N,N,N
OUHCIR-E-8	DOC2VEC	NNNN
STIS-E-1	prop+bert_score+bm25	Y,Y,N,N
STIS-E-2	prop+bert_score	Y,Y,N,N
wut21-E-1	LM Jelinek Mercer	Y,N,N,N

Table 5: Evaluation results of the IR subtask (English).

Run	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-measure
NYUCIN-E-1	0.234	0.246	0.261	0.203	0.261	0.275	0.289
ORGE-E-2	0.191	0.188	0.211	0.199	0.222	0.233	0.248
ORGE-E-7	0.196	0.207	0.209	0.201	0.225	0.244	0.252
STIS-E-1	0.163	0.173	0.202	0.192	0.183	0.200	0.221
STIS-E-2	0.172	0.175	0.201	0.191	0.188	0.201	0.218
ORGE-E-4	0.152	0.163	0.191	0.186	0.177	0.193	0.210
ORGE-E-6	0.152	0.155	0.187	0.185	0.174	0.186	0.206
ORGE-E-3	0.147	0.159	0.187	0.173	0.171	0.192	0.212
ORGE-E-5	0.149	0.158	0.182	0.187	0.175	0.188	0.203
ORGE-E-8	0.144	0.155	0.181	0.175	0.166	0.180	0.194
wut21-E-1	0.176	0.166	0.180	0.101	0.207	0.211	0.226
ORGE-E-1	0.143	0.149	0.179	0.178	0.163	0.175	0.192
OUHCIR-E-5	0.120	0.138	0.153	0.110	0.142	0.161	0.174
OUHCIR-E-4	0.071	0.093	0.126	0.096	0.087	0.107	0.124
OUHCIR-E-6	0.071	0.093	0.126	0.096	0.087	0.107	0.124
OUHCIR-E-3	0.047	0.064	0.083	0.073	0.058	0.075	0.086
KSU-E-9	0.057	0.067	0.069	0.046	0.068	0.080	0.088
KSU-E-7	0.052	0.052	0.051	0.036	0.057	0.063	0.066
KSU-E-5	0.026	0.039	0.044	0.034	0.033	0.046	0.051
KSU-E-1	0.027	0.037	0.039	0.026	0.033	0.044	0.050
KSU-E-3	0.023	0.021	0.028	0.022	0.028	0.030	0.038
OUHCIR-E-1	0.021	0.022	0.025	0.016	0.030	0.032	0.035
OUHCIR-E-2	0.021	0.022	0.025	0.016	0.030	0.032	0.035
OUHCIR-E-8	0.012	0.013	0.019	0.011	0.013	0.017	0.022
OUHCIR-E-7	0.011	0.011	0.012	0.012	0.009	0.009	0.010

Table 6: Evaluation results of the IR subtask (Japanese).

	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-measure
Run							
ORGJ-J-2	0.411	0.424	0.438	0.326	0.454	0.475	0.486
ORGJ-J-7	0.404	0.416	0.434	0.331	0.446	0.465	0.478
ORGJ-J-1	0.415	0.415	0.429	0.317	0.457	0.470	0.480
ORGJ-J-3	0.387	0.402	0.409	0.314	0.427	0.449	0.457
ORGJ-J-5	0.361	0.379	0.405	0.285	0.398	0.424	0.436
ORGJ-J-6	0.381	0.380	0.405	0.321	0.415	0.428	0.443
ORGJ-J-4	0.367	0.365	0.396	0.320	0.408	0.420	0.435
ORGJ-J-8	0.342	0.362	0.385	0.281	0.381	0.405	0.418
KSU-J-10	0.302	0.294	0.314	0.228	0.337	0.349	0.365
UHGSIS-J-10	0.237	0.241	0.260	0.186	0.257	0.268	0.279
UHGSIS-J-2	0.237	0.241	0.260	0.186	0.257	0.268	0.279
UHGSIS-J-4	0.237	0.241	0.260	0.186	0.257	0.268	0.279
UHGSIS-J-6	0.237	0.241	0.260	0.186	0.257	0.268	0.279
UHGSIS-J-8	0.237	0.241	0.260	0.186	0.257	0.268	0.279
UHGSIS-J-1	0.213	0.220	0.234	0.164	0.230	0.243	0.252
UHGSIS-J-3	0.213	0.220	0.234	0.164	0.230	0.243	0.252
UHGSIS-J-5	0.213	0.220	0.234	0.164	0.230	0.243	0.252
UHGSIS-J-7	0.213	0.220	0.234	0.164	0.230	0.243	0.252
UHGSIS-J-9	0.213	0.220	0.234	0.164	0.230	0.243	0.252
KSU-J-2	0.195	0.208	0.218	0.125	0.226	0.247	0.263
KSU-J-4	0.195	0.208	0.218	0.125	0.226	0.247	0.263
KSU-J-8	0.126	0.139	0.151	0.087	0.146	0.165	0.177
KSU-J-6	0.126	0.139	0.151	0.087	0.146	0.165	0.177