A Test Collection for Ad-hoc Dataset Retrieval

Makoto P. Kato
University of Tsukuba / JST, PRESTO
mpkato@acm.org

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

Ying-Hsang Liu
Oslo Metropolitan University
yinghsan@oslomet.no

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

ABSTRACT

This paper introduces a new test collection for ad-hoc dataset retrieval, which have been developed through a shared task called *Data Search* in the fifteenth NTCIR. This test collection consists of dataset collections derived from the US and Japanese governments' open data sites (i.e., Data.gov and e-Stat), as well as English and Japanese topics for these collections. Organizing the shared task in NTCIR, we conducted relevance judgments for datasets retrieved by 74 search systems, and included them in the test collection. In addition to the detailed description of the test collection, we conducted in-depth analysis on the test collection, and revealed (1) what techniques were used and effective, (2) what topics were difficult, and (3) large topic variability in the dataset retrieval task.

CCS CONCEPTS

• **Information systems → Test collections**.

KEYWORDS

Dataset search; ad-hoc retrieval; test collection

ACM Reference Format:
Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2021. A Test Collection for Ad-hoc Dataset Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463261

1 INTRODUCTION

The open data movement is now being accelerated by the expectation for open science and citizen science. Researchers worldwide could collaborate on social problems, and citizens could also participate in research activities if various kinds of data were publicly available. The government of each country has strongly encouraged the open data movement and launched open-data government initiatives such as Data.gov[1] in the United States, Data.gov.uk in the United Kingdom, Data.gov.au in Australia, and e-Stat[2] in Japan. Besides the governmental portal sites, there are also thousands of data repositories on the Web [10].

The open data movement's growth has naturally motivated researchers and industries to develop search engines for the open data scattering on the Web. Google launched Google Dataset Search as a public beta in September 2018 [15], and some researchers have started to discuss potential research topics of dataset search [2]. Although there have been several attempts for understanding and developing dataset search, either a benchmark or an evaluation campaign on dataset search has not been proposed yet.

Therefore, following rapidly increasing demands and interests in dataset search, we organized a shared task, *NTCIR-15 Data Search*[3], at the fifteenth NTCIR [6], and developed a new test collection for ad-hoc dataset retrieval. While the task was named *data search*, a more appropriate name may be *dataset search*, which was defined by Chapman et al. [2]: discovery, exploration, and return of datasets to an end-user, where a dataset is defined as a collection of related observations organized and formatted for a particular purpose. Following the definition given by Chapman et al., we refer to our task as *dataset search* throughout this paper, although the task was named data search.

We aimed to address the following research problems in dataset search through this shared task:

Query understanding for dataset search According to the query log analysis of open data portal sites [5], queries for dataset search include more geographical, temporal, and numerical keywords than those for Web search. Furthermore, as suggested by Koesten *et al.* [8], the goal of dataset search can be diverse, e.g., time series analysis and summarization. Thus, queries for dataset search need a dedicated interpretation technique and to be studied for a better retrieval performance.

Automatic interpretation of dataset contents Metadata of data usually include the name, short description, category, and date. They are used for indexing datasets but are not always sufficiently informative for dataset search. Datasets are often released in Excel, CSV, XML, and PDF formats, and structured in tables or described by RDF for many cases. They could be potentially used together with metadata to enrich the index for dataset search, while interpreting diverse datasets on the Web is still challenging.

Retrieval models for dataset search Datasets and their metadata contain many entities such as location names, temporal expressions, and numerical expressions. Hence, retrieval

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

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463261

models for entity or temporal information could be effective in dataset search as well. Numerical expressions might require a new model for better rankings.

The first round of Data Search focused on retrieval from statistical data collections published by the US government (Data.gov) and Japanese government (e-Stat). We developed a set of topics derived from questions in a community question-answering service and queries through a crowd-sourcing service. Six research groups participated in the NTCIR-15 Data Search task and submitted 74 runs in total. The top-ranked datasets in these runs were then pooled and evaluated by human assessors. Analyzing the submitted runs and conducting per-topic analysis, we found that (1) some approaches are potentially effective to improve the query-metadata matching, (2) topics with location, time, and number expressions are especially difficult, and (3) there were large variability in the topic difficulty and large per-topic system variability in the dataset retrieval task.

The remainder of this paper is organized as follows. Section 2 explains related work on dataset search. Section 3 introduces the task, test collection, and evaluation methodology. Section 4 shows the evaluation results and discusses findings from the analysis. Finally, Section 5 concludes this paper with future directions.

2 RELATED WORK

The initial research on dataset search mainly focused on how people search for datasets. Megler and Maier studied a similarity-based dataset retrieval system by mainly focusing on users' perceptions about dataset similarity, dataset relevance, and users' satisfaction with the proposed system [11, 12]. The keyword-based retrieval systems or search algorithms were not in the scope of their work. Kacprzak *et al.* investigated queries for four national open data portal sites [5]. Their analysis revealed that (1) 90% of queries are 1-3 words queries and the average length is 2.03, (2) location keywords, temporal keywords, file and dataset types, and numbers are included in 5-8% of queries, and (3) there are a small number of question type queries (less than 1%). Koesten *et al.* studied the information seeking behavior in dataset search [8]. Based on interviews with several types of dataset users, they developed a taxonomy of activities with datasets, identified primary relevance criteria such as relevance, usability, and quality, and found a typical workflow after finding relevant datasets. Kern and Mathiak studied how social science researchers actually search for datasets and found several differences from well-known literature search [7]. One of the major findings is that the dataset selection was conducted more carefully than the literature selection. Gregory et al. [4] reviewed scientific articles in five domains and identified key similarities in dataset retrieval practices.

There are several research directions that do not explicitly refer to dataset search but are potentially related. Table search is one of the most related work to dataset search, since datasets are often represented in the form of a table. Zhang and Balog tackled a problem of ad-hoc table retrieval [23], and proposed table-specific features effective for table retrieval and similarity measures for the query-table matching. Table explanation is also a related topic to dataset search since it can be used to enrich indexing or snippet generation in dataset search [1, 9, 18, 20, 21].

Table 1: Statistics of the test collection.

Resource		English	Japanese
Topics	Training topics	96	96
	Test topics	96	96
Collections	Datasets	46,615	1,338,402
	Data files	92,930	1,338,402
Qrels	Training qrels	2,008	2,035
	0: Irrelevant	975 (48.6%)	1,046 (51.4%)
	1: Partially relevant	925 (46.1%)	700 (34.4%)
	2: Highly relevant	108 (5.38%)	289 (14.2%)
	Test qrels	8,528	8,924
	0: Irrelevant	7,986 (93.6%)	5,385 (60.3%)
	1: Partially relevant	509 (5.97%)	2,465 (27.6%)
	2: Highly relevant	33 (0.39%)	1,074 (12.0%)

3 TEST COLLECTION

This section explains the details of the test collection, including the task, topics, queries, dataset collections from which datasets are retrieved, relevance judgments, and evaluation methodology. The statistics of the test collection are shown in Table 1. The test collection is publicly available at https://ntcir.datasearch.jp/.

3.1 Task

The task of Data Search is almost the same as the standard ad-hoc retrieval task and is defined as follows: Given a query for dataset search, a system is expected to return a ranked list of *datasets*. As we introduced earlier, a dataset is conceptually defined as a collection of related observations organized and formatted for a particular purpose. In data portal websites such as Data.gov and e-Stat, multiple related data files form a single dataset together with their metadata. Thus, we operationally define a dataset as a pair of metadata and data files, and use it as a unit of retrieval in this task.

An example of an English dataset is shown in Figure 1. This dataset consists of metadata including "id", "title", and "description", as well as multiple data files in CSV and RDF formats.

3.2 Information Needs

We developed topics by mining real information needs from questions in a community question-answering service. The advantage of this approach over query log mining is the availability of actual needs expressed by texts, as they are usually unclear only with query strings. We first retrieved 3,218 question-answers pairs containing links to the Japanese government open data site, e-Stat, from a Japanese community question-answering service, Yahoo! Chiebukuro[4]. We manually examined each question and extracted 192 questions that indicate information needs for dataset search. They were split to form 96 training topics and 96 test topics.

We manually translated these Japanese information needs into English for developing English information needs. Since Japanese-specific named entities are included in the original needs, they were replaced with counterparts in the US. For example, "Tokyo" was replaced with "New York" and "Japanese mountain yam" was

[4] https://chiebukuro.yahoo.co.jp/

Table 2: Examples of the topics.

Topic ID	Information need	Query
DS1-E-0007	Are there many people who can't drive large trailers?	people can't drive large trailers
DS1-E-0009	How many people have a second house?	many people second house
DS1-E-0014	Which city has a population of about 300,000?	city population 300,000
DS1-E-0033	How many restaurants are there in NY?	many restaurants ny
DS1-E-0060	Which area is the largest producer of potatoes?	largest potato producer

```
{
    'id': '002ece58-9603-43f1-8e2e-54e3d9649e84',
    'title': 'Urban Environment & Transit 2010',
    'description': 'Baltimore City is home to many ...',
    'url': 'https://catalog.data.gov/dataset/002ece58
        -9603-43f1-8e2e-54e3d9649e84',
    'attribution': 'Urban Environment & Transit 2010 (
        https://catalog.data.gov/dataset/002ece58
        -9603-43f1-8e2e-54e3d9649e84) is licensed under
         CC BY 3.0',
    'data': [{'data_filename': '...',
              'data_format': 'csv',
              'data_organization': 'City of Baltimore',
              'data_url': 'https://data.baltimorecity.
                  gov/api/views/gsze-vqaj/rows.csv'},
             {'data_filename': '...',
              'data_format': 'rdf',
              'data_organization': 'City of Baltimore',
              'data_url': 'https://data.baltimorecity.
                  gov/api/views/gsze-vqaj/rows.rdf'},
             ...],
    'data_fields': {
              'Catalog Describedby': 'https://project-
                  open-data.cio.gov/v1.1/schema/catalog
                  .json',
              'Category': 'Neighborhoods',
              'Data First Published': '2014-04-04', ...}
}
```

Figure 1: Example of metadata of an English dataset.

replaced with "potato". Some examples of the translated information needs are shown in Table 2.

3.3 Queries

Since it is not apparent how information needs can be translated into queries, we prepared queries by asking crowd-sourcing workers to input keyword queries based on the presented information needs. A Japanese crowd-sourcing service, Lancers[5], was used for the Japanese topics, while Amazon Mechanical Turk[6] was used to recruit people in the US for the English topics. For each topic, ten workers were given an information need and asked to input a query for dataset search. The exact instruction we provided is:

> You are given a request or a question from someone who wants to get certain information or an answer to the question. Please type some keywords for a web search to provide her/his desired information or answer.

where "a request or a question" refers to an information need, and input keywords were regarded as queries from the user. The request/question and search box were separately placed so that workers cannot input queries by simply selecting some keywords from the request/question.

Table 3: Classification of the information needs and queries.

		Need	Query	
			English	Japanese
Location		**73**	**75**	**54**
	Country	58	63	40
	Region	6	4	7
	City	10	9	7
Time		**35**	**19**	**13**
	Current	15	0	0
	Past time point	12	9	10
	Past period	10	10	3
Number		**33**	**15**	**18**
	Unit	10	1	4
	Age	10	10	7
	Time length	7	0	3
	Amount	8	4	4
	Percentage	2	0	0

We then selected the most representative query for each topic as follows. For each query, we compute the cross entropy between the language models of the topic and query:

$$H(t, q) = -\sum_{w \in q} P(w|q) \log P(w|t) \tag{1}$$

where $P(w|t)$ is estimated by the frequency of w in the queries given for topic t, and $P(w|q)$ is estimated by the frequency of w in query q. Low entropy indicates the closeness of the two language models, suggesting that the query language model is close to that for the entire query set for a topic. This can be considered representativeness in the given topic. Thus, we chose query q that minimizes $H(t, q)$ as the most representative query for topic t. Some examples are shown in Table 2, together with their information needs.

In order to provide a better idea about the developed topics, we classified the information needs and queries. Following an earlier study by Kacprzak *et al.* [5], we first judged if information needs and queries contain the location, time, dataset type, and number expressions. Since we could not find dataset type expressions, we excluded this type from our classification. The other expressions were further categorized into fine-grained classes. Table 3 shows the classification result of 192 information needs and queries[7]. Location

[5] https://www.lancers.jp/

[6] https://www.mturk.com/

[7] A single need or query can be classified into multiple classes, and the sum of the numbers of all the classes is not necessarily equal to their total numbers.

Table 4: Statistics of the dataset collections.

Data.gov		e-Stat	
Data format[8]			
PDF	47,508 (50.9%)	Excel	721,236 (53.9%)
XML	32,726 (35.1%)	CSV	568,042 (42.4%)
CSV	3,983 (4.27%)	PDF	49,124 (3.67%)
Text	3,507 (3.76%)		
JSON	1,946 (2.08%)		
Excel	1,519 (1.63%)		
RDF	1,484 (1.59%)		
others	694 (.743%)		
Licenses			
U.S. Government Works	43,822 (94.0%)	CC BY 4.0	1,338,402 (100%)
CC0 1.0 Universal	1,787 (3.83%)		
PDDL v1.0	512 (1.10%)		
CC BY 3.0	282 (.605%)		
CC BY 4.0	193 (.414%)		
Public Domain Mark 1.0	14 (.030%)		
ODC-By v1.0	5 (.011%)		

expressions are frequently used in both information needs and queries: 38% of the information needs, 39% of the English queries, and 28% of the Japanese queries include either a country, region, or city name. The most frequent location name is “US” for the English queries and “Japan” for the Japanese queries. It is natural since the information needs were obtained from questions related to e-Stat, and later translated into English with “Japan” replaced with “US”. As can be seen from the statistics, some English queries included “US” even though their information need only contains “states”. Time and number expressions are included in about 18% of the information needs and 8% of the queries. Temporal expressions indicating the current time (e.g., “now” and “recent”) were excluded in the queries. This phenomenon could be explained by a hypothesis that users assume a search engine returns recent statistics even though they do not explicitly specify it. In contrast, most of the past time points (e.g., 2010) and past periods (e.g., 1970 – 1990) were also included in the queries, though periods were often omitted in the Japanese queries. Among the number expressions, age expressions (e.g., 20 years old) were usually included in the queries, probably because demographics of statistical data are generally essential factors that distinguish relevant and irrelevant datasets.

3.4 Datasets

We crawled around 0.2 million pages in Data.gov and 1.3 million pages in e-Stat. Each page in these websites describes a single dataset consisting of metadata and data files. While Data.gov datasets include multiple data files, e-Stat datasets include only a single data file. In this round of the Data Search task, for convenience of data processing, we restrict the type of data files to Excel (i.e., xls and xlsx), CSV, and PDF files for e-Stat data files, and Excel, CSV, PDF, XML, JSON, RDF, and text files for Data.gov data files. To increase the availability of the datasets, we used only the datasets allowing redistribution and modification. All the datasets in e-Stats are distributed under a license compatible to CC BY 4.0[9], which allows redistribution and modification. For the Data.gov datasets, we used only the datasets distributed under U.S. Government Work, CC BY, CC0, Public Domain, and Open Data Commons licenses. The statistics of the datasets can be found in Table 4.

Data.gov is a portal site of the US Government’s open data on agriculture, climate, ecosystems, energy, local government, maritime, ocean, and older adults’ health. The metadata consist of the name, ID, short description, category, publishing organization, survey date, and release date. Compared to e-Stat, a little more detailed metadata are given to a set of data files. e-Stat provides diverse kinds of statistical data on weather, population, industry, energy, transportation, education, science, government, and so on.

3.5 Relevance Judgments

Our test collection contains relevance judgments for training and test topics. Relevance judgments for training topics were released to the NTCIR-15 Data Search participants to support their system development. We developed several standard baseline systems such as BM25, LM, and BM25 with RM3, which were implemented by Anserini [22][10], and pooled the top-ranked results for the training topics. The topic-dataset pairs were then evaluated as explained in the next subsection.

Relevance judgments for test topics were conducted with system results submitted by six research groups including two organizer teams. In NTCIR-15 Data Search, there were two types of runs, namely, official runs and extra runs. Official runs are those submitted by the official deadline. Extra runs are those submitted after the official deadline, given an additional call for run submissions from the task organizers. Although the official overview paper [6] only describes official runs, this paper reports the developed test collection based on both of the runs. In total, NTCIR-15 Data Search conducted relevance judgments for 74 runs, of which 43 were for English test topics, and 31 were for Japanese test topics. These runs were pooled in the same way as the relevance judgments for training topics, and evaluated as explained in the next subsection. As a result, the top 10 documents of each system for each query were all judged. The statistics are shown in Table 1.

3.6 Evaluation Methodology

The evaluation of Data Search is almost the same as the standard ad-hoc retrieval evaluation. For both the training and test topics, we pooled the top 10 documents of each system for each topic. The crowd-sourcing services used for the query generation were also used for relevance judgments. Each topic-dataset pair was evaluated at a three-point scale (0: irrelevant, 1: partially relevant, and 2: highly relevant). We presented the webpage explaining the datasets and asked the workers to provide a grade based on the page content. The exact instruction we provided is:

- Please judge how useful a DATASET of a webpage is for answering a given REQUEST.
- Please carefully read a given REQUEST, visit a webpage describing a DATASET, and give a usefulness score (0, 1, or 2) to each of the datasets.

[8] The total number of data files in this table and Table 1 is slightly different as different file formats are assigned to the same file in the metadata of Data.gov.

[9] https://creativecommons.org/licenses/by/4.0/legalcode

[10] The code is available at https://github.com/mpkato/ntcir-datasearch

For training topics, we assigned five workers to each topic-dataset pair and removed the highest and lowest scores for excluding outliers. To ensure the assessments' quality, we showed the same topic-dataset pairs and measured the consistency of the assessments. If over 25% of answers for these topic-dataset pairs were inconsistent, we excluded such assessors in the evaluation. The inter-rater agreement measured by Krippendorff's α is 0.344 for the English topics and 0.736 for the Japanese topics.

Since we found a low agreement for the English topics, we updated the crowd-sourcing settings for test topics as follows. We selected topic-dataset pairs for which relevance scores were highly consistent, and used them as *gold* datasets for measuring the performance of each worker. More precisely, they are considered highly consistent if the average score of *five* scores is 1.8 or higher, or 0.2 or lower. In the evaluation with the test topics, 10% of topic-dataset pairs were used as gold datasets. We banned workers who conducted over 30 judgments and made errors for over 30% of gold datasets. Moreover, we used the option "Require that Workers be Masters to do your tasks" in Amazon Mechanical Turk and found that this setting significantly increased the quality of the judgments. Five assessors were assigned for each topic-dataset pair for the Japanese topics, while three assessors were assigned for the English topics. Finally, we achieved Krippendorff's α of 0.478 for the Japanese topics and 0.438 for the English topics.

Standard evaluation metrics for ad-hoc retrieval tasks, nDCG, ERR, and Q-measure, were used in NTCIR-15 Data Search. nDCG@10 was used as the primary metric in our task, based on an assumption that a single SERP contains ten datasets and users are likely to have informational intents in dataset search [17]. NTCIREVAL was used for computing the effectiveness scores[11].

To increase the reproducibility of the evaluation, we released all the evaluation scripts used in NTCIR-15 Data Search, which are available at https://github.com/mpkato/ntcir-datasearch-evalscripts. They include scripts for pooling, preparing CSV files for crowd-sourcing, and HTML files used in the crowd-sourcing tasks.

4 ANALYSIS

We report the results of in-depth analysis on the developed test collection, and provide insight into the ad-hoc dataset retrieval task.

4.1 System Effectiveness

Figure 2 shows the effectiveness of the runs submitted at NTCIR-15 Data Search, which are grouped by research groups: ORG (organizers' baselines), KSU [16], NII [14], STIS [19], and uhai [13]. The alphabets above the bars indicate techniques used in those runs:

- **T** Table headers. As many of the datasets are represented by tables and their headers are more informative than the metadata, some systems extracted table headers from data files and computed a BM25 score between a query and metadata plus the extracted headers.
- **E** Entity. According to the analysis conducted by Nguyen et al. [14], 99% of the datasets in the Data.gov collection contain entities and 82% of them contain location or time information. Entities were extracted to create a special field, by which an entity-oriented matching score was computed.
- **N** Neural language models including word embedding, BERT, and RoBERTa. These models were mainly used to embed a query and metadata of the dataset and to predict their relevance.
- **C** Category. KSU [16] categorized queries and metadata of the datasets and computed a matching score based on the category overlap. Categories used in a community question-answering service were employed in their systems.
- **Q** Query modification. Queries in the test collection include some terms that hurt the retrieval effectiveness. Such terms were detected and excluded by uhai [13].
- **L** Learning to rank. A standard learning to rank approach was applied with matching scores obtained from the baseline methods and those obtained by the neural language models.

Table 5: The mean nDCG@10 for each need and query type.

		English	Japanese
All		0.192	0.350
Need	Location	0.165	0.327
	Time	0.085	0.297
	Number	0.125	0.244
Query	Location	0.178	0.380
	Time	0.097	0.270
	Number	0.088	0.327

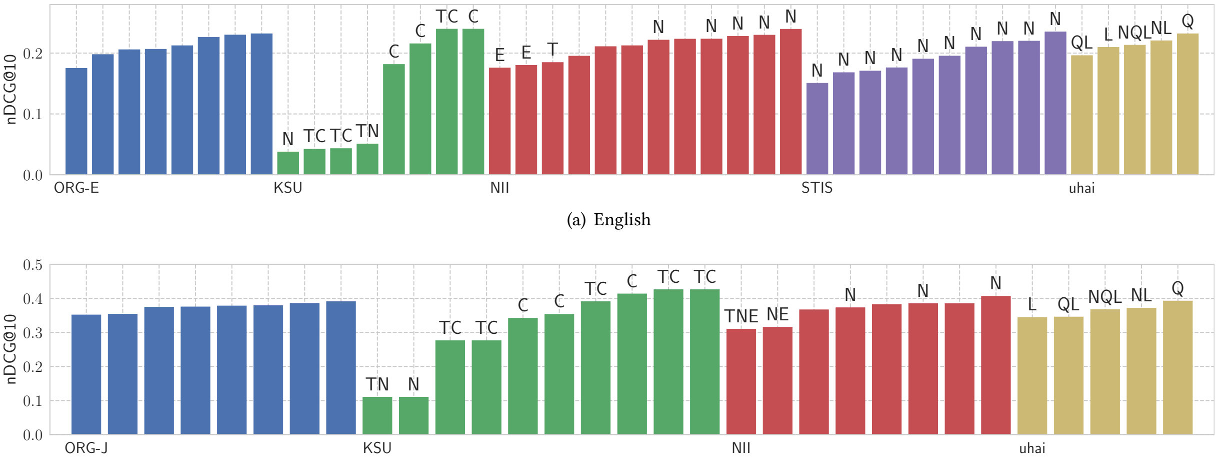
The baseline runs indicated by "ORG-E" and "ORG-J" are standard baseline search models: BM25, BM25 with the pseudo relevance feedback, query likelihood model, and sequential dependency model (see the baseline code for details). These baseline and participants' approaches are mainly based on the similarity between a query and metadata of a dataset, and used the data files for improving the query-metadata matching in some systems.

While there is no clear trend, there are several implications from those evaluation results. First, some approaches including those using the table header (T), categories (C), and neural language models (N) are potentially effective to improve the query-metadata matching in the ad-hoc dataset retrieval task. Second, the nDCG scores are lower for the English topics than those for the Japanese topics. This result could be explained by the difference in the relevance grade distributions: as Table 1 shows, there are only 509 partially relevant and 33 highly relevant documents for the English topics, while there are 2,465 partially relevant and 1,074 highly relevant documents for the Japanese topics. This difference may be caused by the topic development procedure in which we constructed information needs for Japanese topics and later translated them into English. Finally, we note that there is much room for improvement in this task: the ideal system performance, which is the average of the maximum nDCG@10 scores per query that were achieved by the submitted runs, was 0.567 against the highest nDCG@10 score of 0.240 for the English test topics, and 0.731 against the highest nDCG@10 score of 0.426 for the Japanese test topics.

4.2 Topic Difficulty

Table 5 shows the mean nDCG@10 for each information need and query type. "All" row indicates the mean nDCG@10 of all the English or Japanese topics, while "Need" and "Query" rows indicate

[11] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html



(a) English

(b) Japanese

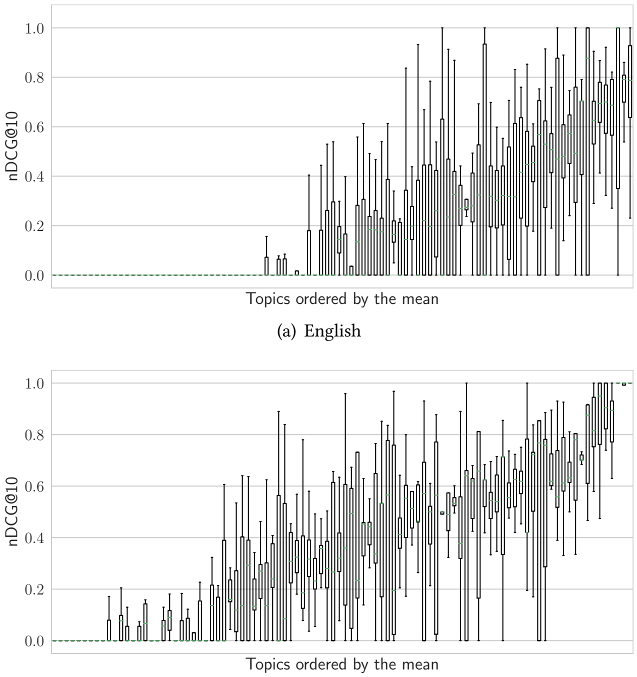
Figure 2: Evaluation results of the runs submitted at NTCIR-15 Data Search.

that of only topics in which the information need or query contains location, time, or number expressions (see Section 3.3 for details).

Lower performances in "Need" and "Query" rows than those in "All" rows suggest that the topics with such unique expressions in the information need or query seem more difficult than the other topics. Among the unique expressions in dataset search, temporal and number expressions are especially difficult compared to location expressions. For example, time ranges (e.g., "between 1868 and 1912") are included in some topics and are hard to retrieve appropriate datasets. These trends are consistent in both languages.

4.3 Topic Variability

Figure 3 visualizes the topic variability in nDCG@10 across all the submitted runs. Boxplots represent the system performance distribution in each topic and are ordered by the mean nDCG@10. Two types of large variability can be seen in this figure. First, the mean nDCG@10 for each topic greatly varies across the topics. There are some queries for which most of the systems could find relevant datasets, while there are also many difficult topics especially for the Data.gov dataset collection. Second, the per-topic variability of the system performances is also large compared to the other retrieval tasks (e.g., TREC Web track [3]). Since there were relatively small differences in the overall system performances, large per-topic variability suggests that no single system performed better than the others for all the topics.



(a) English

(b) Japanese

Figure 3: Topic variability in nDCG@10 across all the submitted runs.

5 CONCLUSIONS

This paper introduced a new test collection for ad-hoc dataset retrieval. In addition to the detailed description of the test collection, in-depth analysis on the developed test collection demonstrated (1) some approaches are potentially effective to improve the query-metadata matching, (2) topics with location, time, and number expressions are especially difficult, and (3) there were large topic variability in terms of the difficulty and large per-topic system variability in the ad-hoc dataset retrieval task.

Building on the success at the first round, we plan to organize the next round of the Data Search task at NTCIR-16, where a new task, question-answering for a dataset collection, will be introduced. 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.

ACKNOWLEDGMENTS

We would like to thank the NTCIR-15 Data Search participants for their contributions to the test collection development. This work was supported by JSPS KAKENHI Grant Numbers JP18H03243 and JP18H03244, and JST PRESTO Grant Number JPMJPR1853, Japan.

REFERENCES

[1] Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *AAAI*. 5020–5027.

[2] 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

[3] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview.* Technical Report.

[4] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2019. Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology* 70, 5 (2019), 419–432.

[5] Emilia Kacprzak, Laura M Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *International Conference on Web Engineering*. 429–436.

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

[7] Dagmar Kern and Brigitte Mathiak. 2015. Are there any differences in data set retrieval compared to well-known literature retrieval?. In *TPDL*. 197–208.

[8] Laura M Koesten, Emilia Kacprzak, Jenifer FA Tennison, and Elena Simperl. 2017. The trials and tribulations of working with structured data:-a study on information seeking behaviour. In *CHI*. 1277–1289.

[9] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *EMNLP*. 1203–1213.

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

[11] V. M. Megler and David Maier. 2015. Are Data Sets Like Documents?: Evaluating Similarity-Based Ranked Search over Scientific Data. *IEEE Trans. Knowl. Data Eng.* 27, 1 (2015), 32–45.

[12] Veronika Margaret Megler and David Maier. 2015. Demonstrating "Data Near Here" Scientific Data Search. In *SIGMOD*. 1075–1080.

[13] Ryota Mibayashi, Pham HuuLong, Naoaki Matsumoto, Takehiro Yamamoto, and Hiroaki Ohshima. 2020. Uhai at the NTCIR-15 Data Search Task. In *Proceedings of the NTCIR-15 Conference*.

[14] Phuc Nguyen, Kazutoshi Shinoda, Taku Sakamoto, Diana Andreea Petrescu, Hung Nghiep Tran, Atsuhiro Takasu, Akiko Aizawa, and Hideaki Takeda. 2020. NII Table Linker at the NTCIR-15 Data Search Task. In *Proceedings of the NTCIR-15 Conference*.

[15] 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.

[16] Taku Okamoto and Hisashi Miyamori. 2020. KSU Systems at the NTCIR-15 Data Search Task. In *Proceedings of the NTCIR-15 Conference*.

[17] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *WWW*. 499–508.

[18] Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. In *AAAI*. 5414–5421.

[19] Lya Hulliyyatus Suadaa, Lutfi Rahmatuti Maghfiroh, Isfan Nur Fauzi, and Siti Mariyah. 2020. STIS at the NTCIR-15 Data Search Task: Document Retrieval Re-ranking. In *Proceedings of the NTCIR-15 Conference*.

[20] Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in Data-to-Document Generation. In *EMNLP*. 2253–2263.

[21] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *SIGMOD*. 97–108.

[22] 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

[23] Shuo Zhang and Krisztian Balog. 2018. Ad Hoc Table Retrieval Using Semantic Similarity. In *WWW*. 1553–1562.