

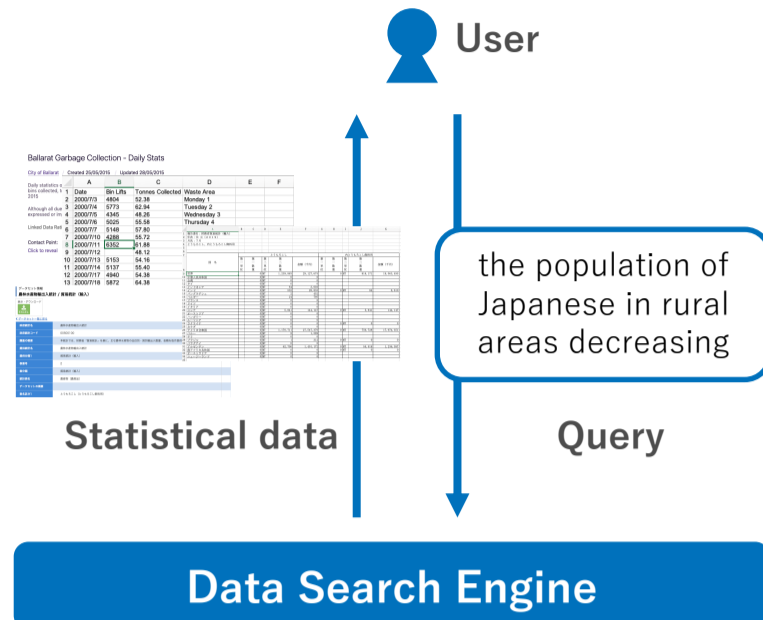
# Overview of the NTCIR-15 Data Search Task

Makoto P. Kato (University of Tsukuba), Hiroaki Ohshima (University of Hyogo), Ying-Hsang Liu (University of Southern Denmark), Hsin-Liang Chen (Missouri University of Science and Technology)

## Task

### The very first IR evaluation campaign for data search

- Subtasks**
  - English and Japanese
- Input**
  - 96 queries for each of the subtasks
- Document (or Dataset) collection**
  - e-Stats for Japanese
  - Data.gov for English
- Output**
  - Ranked list of datasets for each query



## Test Collection Stats.

Language	Documents (or datasets)	Training queries	Test queries	Relevance judgments for training queries	Relevance judgments for test queries	
Japanese	1,338,402	96	96	2,035	5,719	
	English	46,615	96	96	2,008	6,240

## Topics & Queries

- Information needs, by which queries are generated and relevance of a dataset is judged, are derived from questions in cQA**
  - Extracted **3,219** Q&As from Yahoo Chiebukuro (Yahoo Japan's cQA) that include links to a Japanese open data portal
  - They were manually assessed, from which we obtained only **192** questions that can be considered as information needs for datasets
  - Japanese-specific entities were transformed into corresponding US-specific ones
    - e.g. Kansai → East coast, Tokyo → New York

## Dataset Collections

- Japanese**
  - e-Stat
    - <https://www.e-stat.go.jp/>
    - 1,338,402 (~100GB)
- English**
  - Data.gov
    - <https://www.data.gov/>
    - 46,615 (~445GB)

## Examples of topics and queries

Topic ID	Topic	Query
DS1-E-0001	Do people in the East Coast dislike oysters?	oysters dislike east coast
DS1-E-0004	I am looking for evidences of domestic self-sufficiency rate of salt	domestic self salt rate.
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

## Relevance Judgments

- The relevance of each dataset for a given query is judged by crowd-sourcing workers
  - 0: Not-relevant
  - 1: Partially relevant
  - 2: Highly relevant
- Inter-rater agreement
  - Japanese: 0.495
  - English: 0.462 (Not high, but not low in IR evaluation)

**Instructions**

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.

**Rules**

- Carefully read a **REQUEST** (Note: this page contains a few types of requests.)
- Make sure that you visit a webpage that describes a **DATASET**, and judge how useful the **DATASET** is for answering the **REQUEST**.
- Usefulness score is defined as:
  - 0: (Useless) The **DATASET** is not useful to answer the **REQUEST** at all, or was not accessible for some reasons.
  - 1: (Partially useful) The **DATASET** is useful to partially answer the **REQUEST**, but cannot fully answer the **REQUEST**.
  - 2: (Highly useful) The **DATASET** is useful to fully answer the **REQUEST**.

**Cautions**

- You will be rejected if the website is not accessed.
- You will be rejected if the work time is too short.
- There are some **REQUEST** and **DATASET** for which a true usefulness score is known. You will be rejected if your answer is very different from the true answer.
- You will be rejected if your work result has been rejected before.

**REQUEST:** Do people in the East Coast dislike oysters?  
**DATASET:** LINK

0: Useless   1: Partially useful   2: Highly useful

## Evaluation Results

NTCIR-15 Data Search attracted six research groups and received 54 systems' results in total (17 for Japanese and 37 for English)

### Japanese (Top 9 runs)

	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q	Note
KSU-J-5	0.388	0.403	0.448	0.283	0.448	0.477	0.498	BM25 + <b>Category classification</b>
KSU-J-1	0.362	0.381	0.421	0.295	0.423	0.453	0.473	BM25 + Table header + <b>Category classification</b>
ORGJ-J-3	0.407	0.413	0.421	0.325	0.450	0.47	0.484	BM25
uhai-J-10	0.403	0.406	0.415	0.312	0.447	0.466	0.484	BM25 + Query modification
ORGJ-J-2	0.402	0.405	0.415	0.328	0.447	0.467	0.483	BM25 (lucene)
ORGJ-J-6	0.379	0.386	0.406	0.321	0.423	0.447	0.464	Query likelihood
ORGJ-J-1	0.382	0.396	0.405	0.308	0.426	0.452	0.464	BM25 + PRF
ORGJ-J-7	0.380	0.386	0.401	0.323	0.430	0.452	0.471	BM25 + Sequential dependency model
ORGJ-J-4	0.365	0.377	0.400	0.318	0.409	0.433	0.452	Query likelihood + Sequential dependency model

**KSU (Kyoto Sangyo University) achieved the best performances (though there is no significant differences among the tops)**

### Possibly effective techniques

### English (Top 9 runs)

	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q	Note
KSU-E-2	0.204	0.231	0.255	0.238	0.229	0.257	0.276	BM25 + Table header + <b>Category classification</b>
KSU-E-6	0.204	0.231	0.255	0.238	0.229	0.257	0.276	BM25 + <b>Category classification</b>
NII TableLinker-E-4	0.233	0.237	0.248	0.251	0.251	0.264	0.278	BM25 + PRF + <b>BERT Reranking</b>
ORGE-E-2	0.219	0.225	0.238	0.240	0.235	0.250	0.264	BM25 (lucene)
uhai-E-5	0.219	0.225	0.238	0.240	0.235	0.250	0.264	BM25 + Query modification
NII TableLinker-E-10	0.221	0.226	0.237	0.238	0.235	0.248	0.264	BM25 + PRF + <b>BERT Reranking</b>
STIS-E-2	0.23	0.228	0.237	0.217	0.248	0.255	0.264	BM25 + RM3 + <b>BERT Reranking</b>
ORGE-E-7	0.216	0.220	0.236	0.237	0.228	0.242	0.256	BM25 + Sequential dependency model
ORGE-E-8	0.224	0.230	0.233	0.238	0.244	0.255	0.264	Query likelihood + RM3

- Category classifier (used by KSU)**
  - Train a category classifier by cQA datasets and applied it to queries and documents
  - A document is considered relevant if its category is the same as that of a query
  - A simple, but effective technique that can be seen in production systems
- Dataset header (used by KSU and NII)**
  - The headers of datasets were also used as a part of documents
  - Possibly effective but may need more exploration
- BERT (used by all the teams)**
  - A successful technique often used in NLP tasks
  - Not conclusive again, probably due to lack of large training data