NTCIR-16
Data Search 2

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)




- **The open data movement is now being accelerated by the expectations for open science and citizen science**
  - Each country strongly encourages the open data movement:
    - Data.gov (United States)
    - Data.gov.uk (United Kingdom)
    - Data.gov.au (Australia)
    - e-Stat (Japan)
- **Besides the governmental portals, there are also thousands of data repositories on the Web**

Demand for a better data search engine

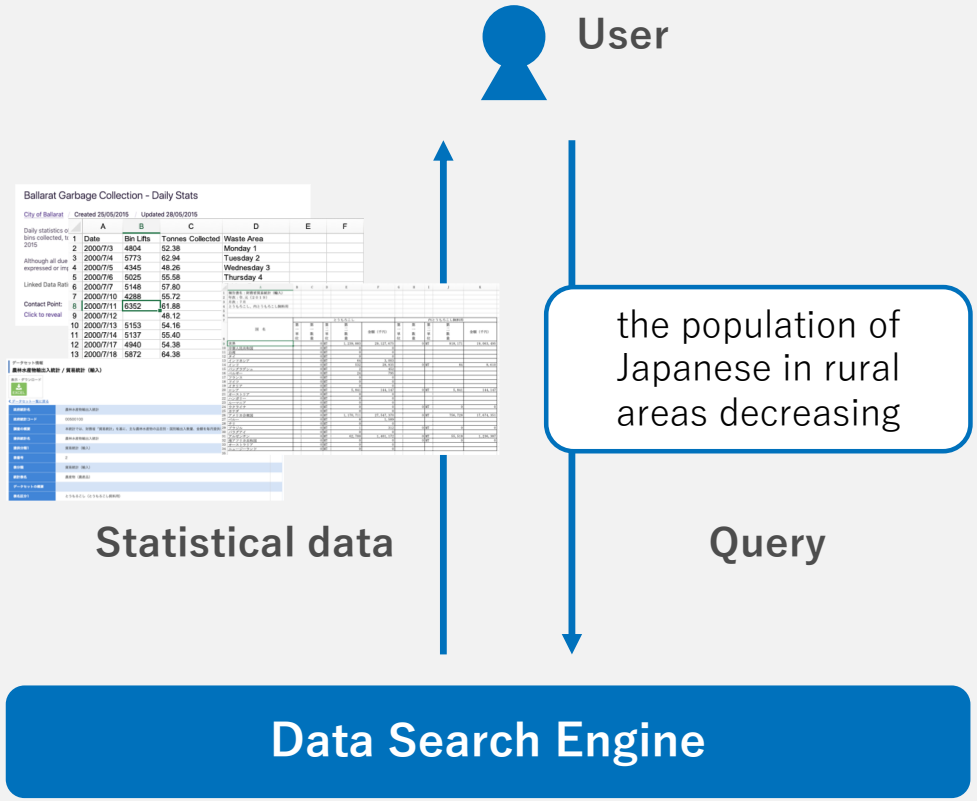
(e.g. Google Dataset Search)

The very first IR evaluation campaign for data search

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

Ad-hoc retrieval for statistical data

- **Subtasks**
  - English and Japanese
- **Input**
  - 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



Examples of Topics

Topic ID	Need	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

Examples of Data

Ballarat Garbage Collection - Daily Stats

City of Ballarat / Created 25/05/2015 / Updated 28/05/2015

Daily statistics of garbage collection in the City of Ballarat. Includes date, number of garbage bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015

Although all due care has been taken to ensure that these data are correct, no warranty is expressed or implied by the City of Ballarat in their use.

Linked Data Rating: ★★☆☆☆

Contact Point:

Click to reveal

	A	B	C	D	E	F
1	Date	Bin Lifts	Tonnes Collected	Waste Area		
2	2000/7/3	4804	52.38	Monday 1		
3	2000/7/4	5773	62.94	Tuesday 2		
4	2000/7/5	4345	48.26	Wednesday 3		
5	2000/7/6	5025	55.58	Thursday 4		
6	2000/7/7	5148	57.80	Friday 5		
7	2000/7/10	4288	55.72	Monday 1		
8	2000/7/11	6352	61.88	Tuesday 2		
9	2000/7/12		48.12	Wednesday 3		
10	2000/7/13	5153	54.16	Thursday 4		
11	2000/7/14	5137	55.40	Friday 5		
12	2000/7/17	4940	54.38	Monday 1		
13	2000/7/18	5872	64.38	Tuesday 2		
14	2000/7/19	4188	47.02	Wednesday 3		
15	2000/7/20	5057	54.26	Thursday 4		
16	2000/7/21	5063	54.38	Friday 5		

データセット情報

農林水産物輸出入統計 / 貿易統計（輸入）

表示・ダウンロード

EXCEL

データセット一覧に戻る

政府統計名	農林水産物輸出入統計
政府統計コード	00500100
調査の概要	本統計では、財務省「貿易統計」を基に、主な農林水産物の品目別・国別輸出入数量、金額を毎月提供しています。
提供統計名	農林水産物輸出入統計
提供分類1	貿易統計（輸入）
表番号	2
表分類	貿易統計（輸入）
統計表名	農産物（農産品）
データセットの概要	
表名区分1	とうもろこし（とうもろこし飼料用）

	A	B	C	D	E	F	G	H	I	J	K
1	報告書名：財務省貿易統計（輸入）										
2	年次：令.元（２０１９）										
3	月次：７月										
4	とうもろこし、内とうもろこし飼料用										
5											
6			とうもろこし				内とうもろこし飼料用				
7	国　名	第一単位	第一数量	第二単位	第二数量	金額（千円）	第一単位	第一数量	第二単位	第二数量	金額（千円）
8											
9	世界		0	MT	1,239,883	29,127,675		0	MT	818,171	19,063,495
10	中華人民共和国		0	MT	0	0					
11	台湾		0	MT	0	0					
12	タイ		0	MT	0	0					
13	インドネシア		0	MT	64	3,003					
14	インド		0	MT	532	28,835		0	MT	84	8,610
15	バングラデシュ		0	MT	2	452					
16	ベルギー		0	MT	24	795					
17	フランス		0	MT	0	0					
18	ドイツ		0	MT	0	0					
19	イタリア		0	MT	0	0					
20	ロシア		0	MT	5,841	144,147		0	MT	5,841	144,147
21	オーストリア		0	MT	0	0					
22	ハンガリー		0	MT	0	0					
23	ルーマニア		0	MT	0	0					
24	ウクライナ		0	MT	0	0		0	MT	0	0
25	カナダ		0	MT	0	0					
26	アメリカ合衆国		0	MT	1,170,711	27,547,370		0	MT	756,728	17,674,351
27	ペルー		0	MT	8	1,589					
28	チリ		0	MT	0	0					
29	ブラジル		0	MT	1	312		0	MT	0	0
30	パラグアイ		0	MT	0	0					
31	アルゼンチン		0	MT	62,700	1,401,172		0	MT	55,518	1,236,387
32	南アフリカ共和国		0	MT	0	0		0	MT	0	0
33	オーストラリア		0	MT	0	0					
34	ニュージーランド		0	MT	0	0					
35											

Results from Data Search "1"

Japanese

	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 + **Category classification**
KSU-J-1	0.362	0.381	0.421	0.295	0.423	0.453	0.473	BM25 + Table header + **Category classification**
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			
ORGJ-J-1	0.382	0.396	0.405	0.308	0.426			
ORGJ-J-7	0.380	0.386	0.401	0.323	0.430			
ORGJ-J-4	0.365	0.377	0.400	0.318	0.409			

Six teams participated in the first round

English

	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 + **Category classification**
KSU-E-6	0.204	0.231	0.255	0.238	0.229	0.257	0.276	BM25 + **Category classification**
NIITableLinker-E-4	0.233	0.237	0.248	0.251	0.251	0.264	0.278	BM25 + PRF + **BERT Reranking**
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
NIITableLinker-E-10	0.221	0.226	0.237	0.238	0.235	0.248	0.264	BM25 + PRF + **BERT Reranking**
STIS-E-2	0.23	0.228	0.237	0.217	0.248	0.255	0.264	BM25 + RM3 + **BERT Reranking**
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

Question Answering Subtask

- **Given a question and a dataset, extract the answer to the question from the dataset**
  - e.g. Which city has a population of about 300,000?
  - e.g. What is the current population in Tokyo?
- **Evaluated by MRR**

Search Interface Subtask

- **Participants are expected to develop a system for data search, compare it with our baseline search system, and share findings at the conference**
- **Will provide five topics for comparison, and a baseline search system**

And new test queries for the ad-hoc retrieval task

Schedule

Jan 31, 2021	Dataset collections and training queries release
Aug 31, 2021	Registration due and test queries release
Sep 30, 2021	Run submission due for the ad-hoc retrieval subtask
Nov 31, 2021	Evaluation results release
Feb 1, 2022	Run submission due for the question answering and search interface subtasks
Feb 1, 2022	Draft task overview paper release

- **See http://ntcir.datasearch.jp for more information**