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)









Introduction

- 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

The very first IR evaluation campaign for data search

Japanese	Documents (or <i>datasets</i>)	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 <i>datasets</i>)	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

-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



Query understanding for data search

- -Queries for data search include more geographical, temporal, and numerical keywords than those for Web search (Kacprzak+ 2017)
- -The goal of data search can be diverse, e.g. time series analysis and summarization (Koesten+ 2017)

Data understanding for data search

- -Metadata are not always sufficiently informative
- Data in Excel, CSV, XML, and PDF formats is potentially used with metadata to enrich the index for data search, while interpreting data on the Web is a still challenging problem

Retrieval models for data search

- Data contains a lot of entities such as locations or products, temporal expressions, and numerical expressions
- -Numerical expressions might require a new model for better rankings

- 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 USspecific ones
 - e.g. Kansai \rightarrow East coast, Tokyo \rightarrow New York

Japanese information needs	English information needs
若者の農業離れが最近騒がれていますが、それを裏付けるデータを 探しています。	I am looking for data proving young people's lack of interest in farming.
関西地方では現在も納豆をあまり食べないのですか?	Do people in the East Coast dislike oysters?
老人医療費が医療費全体に占める割合が高いのは本当でしょうか?	Is it true that medical expenses for elderly account for a large percentage of total medical expenses?

Generating Queries

- Used crowd-sourcing services to convert information needs to queries
 - -Showed a need and asked workers to input a query without looking at the need
 - Tried to simulate a more realistic situation
 - -Selected the most "probable" query from 10 workers' queries
 - Built a unigram language model from those queries, and selected the one with the highest perplexity with respect to the language model

νĽ2-
1べての作業を確認することはできません。プレビューは一部の作業のみ確認する事ができます。
<u>次の作業</u> ▶
検索用キーワードの作成作業
ある情報を知りたがっている人,もしくは,ある疑問を解決したい人の要望が以下に提示されます. その人のために,グーグルやヤフーなどで検索をするための検索キーワードを考えて入力してくださ い.
1. まず 「質問・要望」 をよく読んで理解してください. 2 画面を下の方にスクロールさせると「 総変キーワード入力欄」 がありますので、グーグルやヤフー
などで検索をするための検索キーワードをご自身で考えて入力してください。 決して「質問・要望」を見ながら検索キーワードを入力しないでください、コピー&ペーストも厳 響です。
B)
 表示される「質問・要望」の例:「千葉県では日本一落花生がとれますか?」 通切な「検索キーワード」の例:「千葉県 落花生 生産量」
 表示される「質問・要望」の例:「最近,東京都内ではどれくらいマンションが増えているのでし
ょうか?」 • 適切な 『検索キーワード』 の例:「2019年 都内 マンション 建設数」
注意点
、決して「質問・要望」を見ながら検索キーワードを入力しないでください、コピー&ペーストも厳 葉です。
32数単語を入力する場合には空白(スペース)で区切ってください。 1回分の検索用のキーワードを入力してください、2回に渡って検索することを前提としてないでください。
質問・要望
都道府県別の有効求人倍率をおしえてください。

検索キーワード

Topic ID	Торіс	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

Dataset Collections

Japanese

-e-Stat

- <u>https://www.e-stat.go.jp/</u>
- •1,338,402 (~100GB)

データセット情報

作物統計調査 / 作況調査(水陸稲、麦類、豆類、かんしょ、飼料作物、工芸農 作物) 速報 令和元年産一番茶の摘採面積、生葉収穫量及び荒茶生産量(主産 県)



< データセット一覧に戻る

政府統計名	作物統計調査	0
政府統計コード	00500215	
調査の概要	本調査は、毎年、耕地の状況、収穫量等を調査し、耕地面積、農 作物の作付面積、収穫量、被害面積・被害量等を、全国、都道府 県(主産県)別等に提供しています。	
提供統計名	作物統計調査	
提供分類1	作況調査(水陸稲、麦類、豆類、かんしょ、飼料作物、工芸農作物)	
提供分類2	速報	
提供分類3	令和元年産一番茶の摘採面積、生葉収穫量及び荒茶生産量(主産	

English

- -Data.gov
 - •<u>https://www.data.gov/</u>
 - •46,615 (~445GB)



Examples of Datasets

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: ★ ஹ்ஹ் 🕐

Contact Point:

Click to reveal

	Α	В	С	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 表示・ダウンロード EXCEL <l< th=""><th>計 / 貿易統計(輸入)</th><th></th></l<>	計 / 貿易統計(輸入)	
政府統計名	農林水產物輸出入統計	0
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	農林水產物輸出入統計	
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)	

	А	В	С	D	Е	F	G	Н	Ι	J	К
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											
7					とうもろこし				内	とうもろこし飼料用	
		第	第	第	第		第	第	第	第	
	国名	_		-	-		_	-	-	-	
		畄	浙	鼡	数	金額(千円)	崩	数	鼡	批	金額(千円)
8		位	量	位	品		位	量	位	が 量	
0			244	O MT	1 230 883	20 127 675	100	-	MT	818 171	19 063 495
10	世が	ŧ →		0 MT	1,235,665	25, 121, 010			/ 11 /	010, 171	19,003,495
11	台湾	t i		0 MT	0	0					
12	91	1		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		() 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		() 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			MT	0	0
24	<u>77717</u>			O MT	0	0	-		MI	0	0
20	フナク	+ +		O MT	1 170 711	27 547 270			MT	756 709	17 674 951
27	ブクラル首衆国	+		0 MT	1, 110, 111	21, 041, 370		,	/ 01.1	100,128	11,014,301
28	チリ	+		0 MT	0	1, 009			+		
29	プラジル	1		0 MT	1	312		() MT	0	0
30	パラグアイ	1 1		0 MT	0	012				Ŭ	, i i i i i i i i i i i i i i i i i i i
31	アルゼンチン	1		0 MT	62,700	1,401,172		() MT	55, 518	1,236,387
32	南アフリカ共和国			0 MT	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											

"Dataset" is a unit of retrieval in Data Search

-Consists of metadata and multiple data files

•e-Stat

- A data file for each metadata



• Data.gov

 Multiple data files for each metadata

Metadata Data files E Ballarat Garbage Collection - Daily Stats d Waste Area Tonne 52.38 62.94 48.26 55.58 57.80 55.72 4804 5773 4345 5025 5148 4299 Monday 1 Tuesday 2 Wednesda Thursday 4 Friday 5 Monday 1 3 2000/7/4 5 2000/7/6 6 2000/7/7 7 2000/7/1 City of Ballarat / Created 25/05/2015 / Updated 28/05/2015 Tuesday : 61.88 48.12 54.16 55.40 54.38 64.38 64.38 47.02 54.26 54.38 Daily statistics of garbage collection in the City of Banarat. Includes date, nul Wednesda Thursday Friday 5 9 2000/7/12 10 2000/7/13 5153 11 2000/7/14 5137 12 2000/7/14 5137 12 2000/7/18 5872 14 2000/7/19 4188 15 2000/7/19 5057 16 2000/7/21 5063 bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015 GIBSONIA PITTSBURG NATRONA P NATRONA P NATRONA P Although all due care has been taken to ensure that these data are correct, no warranty is LOMBARD S FORBES J ROOSEVELT J CARNEGIE S BUTLER S KULLY S UNION S MICHARL S MICHARL S HOUNTAIN S PLEASANTY R DAVIS S DAVIS S expressed or implied by the City of Ballarat in their use. Contact Point: Click to reveal

• e-Stats

- -Distributed under a license compatible to CC BY, which allows redistribution and modification
- Data.gov

-Used only the datasets that can be redistributed, i.e. U.S. Government Work, CC BY, etc.

- The Data Search test collection is freely available from our website
 - -https://ntcir.datasearch.jp/

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

- 1. Carefully read a REQUEST (Note: this page contains a few types of requests.)
- 2. Make sure that you visit a webpage that describes a **DATASET**, and judge how useful the **DATASET** is for answering the **REQUEST**.
- 3. 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.
- 1.

REQUEST: Do people in the East Coast dislike oysters?

DATASET: LINK

O 0: Useless O 1: Partially useful O 2: Highly useful

Baseline Methods

Applied standard retrieval models to only the metadata



Baseline retrieval models

- BM25, BM25 + RM3, BM25 + SDM, BM25 + BM25PRF
- Query Likelihood, Query Likelihood + RM3, Query Likelihood + SDM

Tools developed for Data Search

⊒mpkato / ntcir-datasearch			Q	Unwatch -	🖈 Star	0 ¥ Fork 0
<>Code (!) Issues 0 (!) Pull rec	quests 0 🔹 Actions	Projects 0	🗉 Wiki 🕕 Secu	urity 🛄 Insi	ghts 🔅 Set	ttings
Baselines for NTCIR Data Search						Edit
-0- 27 commits	ဖို 1 branch	🗊 0 packages	ି ୦	releases	11 1	I contributor
Branch: master - New pull request			Create new file	Upload files	Find file	Clone or download -
Mpkato add preprocessor for en				🗸 La	itest commit 93	c8bd8 20 days ago
.circleci	bug fix f	for circleci config				20 days ago
🖻 anserini @ a972d33	change	the branch of anseri	ni			2 months ago
baselines	add pre	processor for en				20 days ago
collections	add thre	ee dirs				2 months ago
🖬 data	add thre	ee dirs				2 months ago
indices	add thre	ee dirs				2 months ago
.gitignore	ignore d	locker as well				2 months ago
.gitmodules	add sub	module anserini				2 months ago
README.md	add ntci	irify				20 days ago
poetry.lock	add tqd	m				20 days ago
pyproject.toml	add tqd	m				20 days ago
tasks.py	add pre	processor for en				20 days ago
Baselir	ie meth	ods fo	or Dat	ta Se	earcl	า
	Inased	on Ar	rserin			

📮 mpkato / ntcir-data	search-evalscri	ipts	 Unwatch 	• 1	☆ Star 4	얓 Fork 0
<> Code (!) Issues	ໃງ Pull requests	▹ Actions [Projects	🕮 Wiki	Security	
۶º main -	G	to to file Add	file - ⊻	Code -	About	礅
mpkato first commit			on 10 O	ct 🗓 1	Scripts for rep relevance judg	producing gments in the
README.amt.md	first commit		2 mo	nths ago	task	a Search
B README.lancers.md	first commit		2 mo	nths ago	🛱 Readme	
C README.md	first commit		2 mo	nths ago		
amt-template.html	first commit		2 mo	nths ago	Releases	
C compute_median.py	first commit		2 mo	nths ago	No releases publis	shed
Create_amt_task.py	first commit		2 mo	nths ago	Create a new relea	ase
Create_lancers_task	first commit		2 mo	nths ago	Packages	
data_search_e_gold	first commit		2 mo	nths ago	No packages publ	ished
data_search_e_topi	first commit		2 mo	nths ago	Publish your first	package
data_search_j_gold	first commit		2 mo	nths ago		
data_search_j_topic	first commit		2 mo	nths ago	Languages	

Evaluation scripts including detailed crowd-sourcing settings

 NTCIR-15 Data Search attracted six research groups and received 54 systems' results in total

 17 for Japanese and 37 for English

- Overall, the evaluation results showed that
 - -1. Much room for improvement for data search algorithms
 - -2. No single system significantly outperformed the others

•STIS : Politeknik Statistika

- NII: National Institute of Informatics
- •Uhai: University of Hyogo
- KSU: Kyoto Sangyo University

Team	Run	Description
KSU	KSU-J-1	category search, QA categories and BM25 and table headers
KSU	KSU-J-3	Birch and table headers
KSU	KSU-J-5	category search, QA categories and BM25
KSU	KSU-J-7	Birch
uhai	uhai-J-6	Query Fixing + L2R + Bert
uhai	uhai-J-7	Query Fixing + L2R
uhai	uhai-J-8	L2R
uhai	uhai-J-9	L2R + Bert
uhai	uhai-J-10	Query Fixing + bm25

Team	Run	Description
KSU	KSU-E-2	category search, QA categories , BM25 and table headers
KSU	KSU-E-4	Birch and table headers
KSU	KSU-E-6	category search, QA categories and BM25
KSU	KSU-E-8	Birch
NIITableLinker	NIITableLinker-E-1	BM25 [fine-tune]
NIITableLinker	NIITableLinker-E-2	BM25+PRF [default]
NIITableLinker	NIITableLinker-E-3	BM25+PRF [fine-tune]
NIITableLinker	NIITableLinker-E-4	R2+ BERT
NIITableLinker	NIITableLinker-E-5	R3+ BERT
NIITableLinker	NIITableLinker-E-6	Entity + Noun phrase + BM25+PRF
NIITableLinker	NIITableLinker-E-7	DATE LOC
NIITableLinker	NIITableLinker-E-8	metadata attributes + BM25+PRF
NIITableLinker	NIITableLinker-E-9	cluster
NIITableLinker	NIITableLinker-E-10	R3+BERT+Top100

Team	Run	Description
STIS	STIS-E-1	RM3+BM25 AND FINETUNED BERT BERT-BASE-UNCASED
STIS	STIS-E-2	RM3+BM25 AND FINETUNED BERT BERT-BASE-UNCASED
STIS	STIS-E-3	RM3+BM25 AND FINETUNED BERT BERT-LARGE-UNCASED
STIS	STIS-E-4	RM3+BM25 AND FINETUNED BERT BERT-LARGE-UNCASED
STIS	STIS-E-5	RM3+BM25 AND FINETUNED ROBERTA ROBERTA-BASE
STIS	STIS-E-6	RM3+BM25 AND FINETUNED ROBERTA ROBERTA-BASE
STIS	STIS-E-7	RM3+BM25 AND ENCODER CONCAT GLOVE
STIS	STIS-E-8	RM3+BM25 AND ENCODER CONCAT GLOVE
STIS	STIS-E-9	RM3+BM25 AND ENCODER CONCAT GLOVE
STIS	STIS-E-10	RM3+BM25 AND FINETUNED BERT BERT-BASE-UNCASED
uhai	uhai-E-1	Query Fixing + L2R + Bert
uhai	uhai-E-2	Query Fixing + L2R
uhai	uhai-E-3	L2R + Bert
uhai	uhai-E-4	L2R
uhai	uhai-E-5	Query Fixing + bm25



The blue line indicate the maximum score for each query. Each of the other lines indicates one of the top 3 runs. **No single system achieved satisfactory results.**



The blue line indicate the maximum score for each query. Each of the other lines indicates one of the top 3 runs. **No single system achieved satisfactory results.**



Which Team was Successful?

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	KGH	(Kuch	to Sangua University) is
ORGJ-J-7	0.380	0.386	0.401	0.323	0.430	N3U	(Nyo)	to Sangyo University) is
ORGJ-J-4	0.365	0.377	0.400	0.318	0.409	the to	p per	former in both subtasks

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

Not very conclusive yet, but

- 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

Summary

- The very first IR evaluation campaign for data search
- Ad-hoc retrieval for statistical data
- Evaluation results suggested that
 - -1. Much room for improvement for data search algorithms
 - -2. No single system significantly outperformed the others
- NTCIR-16 Data Search 2 (if accepted)
 - -Ad hoc retrieval task
 - -Question answering
 - -Search interface