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)

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

Evaluation Results

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

Possible effective techniques

- 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

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