CWI

CWI @ TREC 2013: Federated Web Search Track

Alejandro Bellogín, Jiyin He, Arjen P. de Vries

Information Access
CWI

ERCIM

CWI

Summary

- We participated in the two tasks
  - Resource selection
  - Results merging
- Query likelihood alone obtains good results for results merging
- We tested our methods with the FedWeb 2012 collection
  - The performance is not consistent in 2013

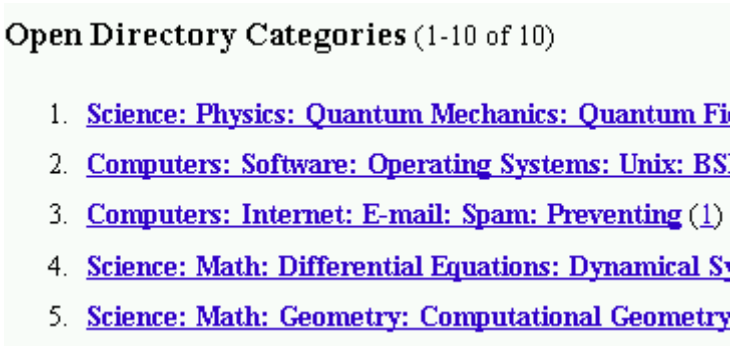
CWI

FedWeb 2012

- 108 resources (vs 156+1 in 2013)
- Top 10 results
  - Snippets
  - Pages
- Crawled in Dec 2011 – Jan 2012 (vs Apr – May 2013)
- TREC queries 51 – 100 (vs 7001 – 7506 ≈ 200 topics)

Resource selection – Methods

- ODP: similarity between ODP's query and resource categories
  - Similarity function: Jaccard vs Cosine
  - Importance of the category according to its ranking?
  - Consider the query text?

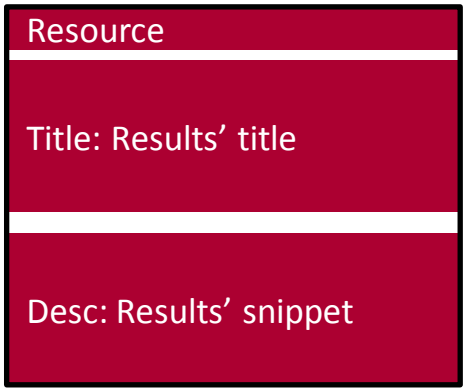

Open Directory Categories (1-10 of 10)

1. Science: Physics: Quantum Mechanics: Quantum Fi
2. Computers: Software: Operating Systems: Unix: BS
3. Computers: Internet: E-mail: Spam: Preventing (1)
4. Science: Math: Differential Equations: Dynamical Sy
5. Science: Math: Geometry: Computational Geometry

Resource selection – Methods

- ODP: similarity between ODP's query and resource categories
  - Similarity function: Jaccard vs Cosine
  - Importance of the category according to its ranking?
  - Consider the query text?

- Retrieval model: build a pseudo-document and retrieve best matching resources
  - Lucene's TF-IDF, BM25, Language models
  - Only title, only snippet, both



Resource selection – Methods

- ODP: similarity between ODP's query and resource categories
  - Similarity function: Jaccard vs Cosine
  - Importance of the category according to its ranking?
  - Consider the query text?

Open Directory Categories (1-10 of 10)

1. Science: Physics: Quantum Mechanics: Quantum Fi
2. Computers: Software: Operating Systems: Unix: BS
3. Computers: Internet: E-mail: Spam: Preventing (1)
4. Science: Math: Differential Equations: Dynamical Sy
5. Science: Math: Geometry: Computational Geometry

- Retrieval model: build a pseudo-document and retrieve best matching resources
  - Lucene's TF-IDF, BM25, Language models
  - Only title, only snippet, both

- Hybrid: Borda voting to aggregate results from the other methods

CWI

Resource selection – Results (2012)

- Tested several variants on FedWeb 2012

Method	MAP	nDCG	MRR
TF-IDF+ODP Jacc	**0.338**	**0.516**	0.564
TF-IDF	0.285	0.412	**0.610**
ODP Jaccard	0.283	0.471	0.439
BM25 (1.2, 0.2)	0.283	0.400	0.545
LM ($\lambda = 0.1$)	0.280	0.407	0.590
ODP Cosine	0.278	0.462	0.400
BM25 (1.2, 0.8)	0.272	0.397	0.557
LM ($\lambda = 0.5$)	0.263	0.394	0.571
LM ($\lambda = 0.9$)	0.252	0.387	0.566
LM ($\lambda = 0.1$) desc	0.241	0.386	0.602
LM ($\mu = 200$)	0.240	0.378	0.551
LM ($\mu = 2000$)	0.240	0.378	0.551
BM25 (1.2, 0.8) desc	0.239	0.383	0.608
TF-IDF title	0.215	0.321	0.495

- Submitted top 3: TF-IDF, ODP Jaccard, and hybrid

CWI

Resource selection – Results (2013)

- Results not consistent in FedWeb 2013 collection

Method	Run	nDCG@20	ERR@20
BM25 (1.2, 0.8) desc	-	**0.1588**	**0.0204**
LM ($\lambda = 0.1$) desc	-	0.1476	**0.0204**
BM25 (1.2, 0.2)	-	0.1346	0.0068
LM ($\lambda = 0.1$)	-	0.1322	0.0068
TF-IDF	CWI13SniTI	0.1235	0.0067
BM25 (1.2, 0.8)	-	0.1223	0.0102
LM ($\lambda = 0.5$)	-	0.1218	0.0051
LM ($\lambda = 0.9$)	-	0.1153	0.0041
LM ($\mu = 2000$)	-	0.1033	0.0051
LM ($\mu = 200$)	-	0.1017	0.0051
TF-IDF title	-	0.1016	0.0017
TF-IDF+ODP Jacc	CWI13ODPTI	0.0961	0.0034
LM ($\lambda = 0.9$)	-	0.0934	0.0017
ODP Jaccard	CWI13ODPJac	0.0497	0.0000

- Not even the fields in the pseudo-document perform the same
  - Best: entire document (2012) vs snippet (2013)

Results merging – Methods

- Based on document relevance
  - Query likelihood (QL)

$$p(d|q) \propto \prod_{w \in q} p(w|d)$$

Results merging – Methods

- Based on document relevance
  - Query likelihood (QL)

$$p(d|q) \propto \prod_{w \in q} p(w|d)$$

- Based on resource selection
  - Cluster: for each ranked resource, the retrieved documents are ranked with QL
  - Diversity: IA-select of QL ranking with respect to the resources

resource

$$P(S|q) = \sum_{c} P(c|q)(1 - \prod_{d \in S}(1 - V(d|q,c)))$$

  - Boost: use the relevance with respect to the resource to boost the documents

$$p(d|q,z) \propto p(d|q)p(q|z)$$

Results merging – Results

Best 2013 resource selection results

Best 2012 resource selection results

Also 2012!

Method	P@10	nDCG@20	nDCG@50	nDCG
2013 data				
CWI13bstBM25desc*	**0.3408**	0.1224	0.2024	0.5366
CWI13IndriQL	0.3220	**0.1622**	**0.2371**	**0.5438**
CWI13iaTODPJ	0.2840	0.1509	0.1915	0.5253
CWI13bstTODPJ	0.2500	0.1466	0.1839	0.4973
CWI13clTODPJ*	0.1940	0.0551	0.0892	0.4610
2012 data				
CWI12bstTODPJ*	**0.4960**	0.1246	0.1989	0.6081
CWI12IndriQL*	0.4900	**0.1464**	**0.2627**	**0.6525**
CWI12clTODPJ*	0.2200	0.0666	0.1106	0.5462
CWI12iaTODPJ*	0.1940	0.0532	0.1015	0.5407

Discussion

- Results merging can be solved with simple IR techniques
  - Query likelihood obtained very good results

- How to define a training set for an evolving test environment?
  - The document rankings of the resources change
  - The content of the websites change
  - The type of queries is important
    - tailored to be answered by a specific resource?
    - time-aware?
    - …