CWI @ TREC 2013: Federated Web Search Track

Alejandro Bellogín, Jiyin He, Arjen P. de Vries

Centrum Wiskunde & Informatica, The Netherlands

{a.bellogin, j.he, arjen.de.vries}@cwi.nl

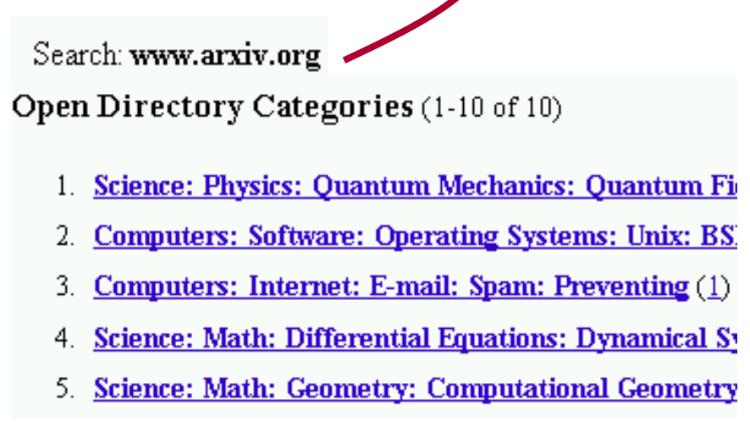
Text REtrieval Conference (TREC)
...to encourage research in information retrieval from large text collections.

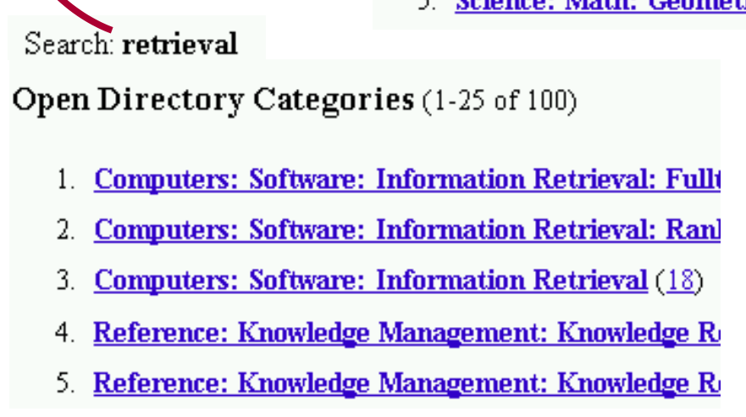
Resource selection

ODP

Similarity between ODP’s query and resource categories

Jaccard
Cosine


Search: www.arxiv.org
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


Search: retrieval
Open Directory Categories (1-25 of 100)
1. Computers: Software: Information Retrieval: Fullt
2. Computers: Software: Information Retrieval: Ranl
3. Computers: Software: Information Retrieval (18)
4. Reference: Knowledge Management: Knowledge R
5. Reference: Knowledge Management: Knowledge R

no order
no query text

Retrieval model

Build pseudo-document and retrieve best matching resources

Resource
Title: Results’ title
Desc: Results’ snippet

TF-IDF
BM25
Language Models

Hybrid run

Aggregates rankings from the other methods using Borda voting

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

Best: Hybrid, TF-IDF, and ODP with Jaccard
Expected similar performance in 2013…

FedWeb 2013

Submitted best methods on 2012 collection

Results do not agree with 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

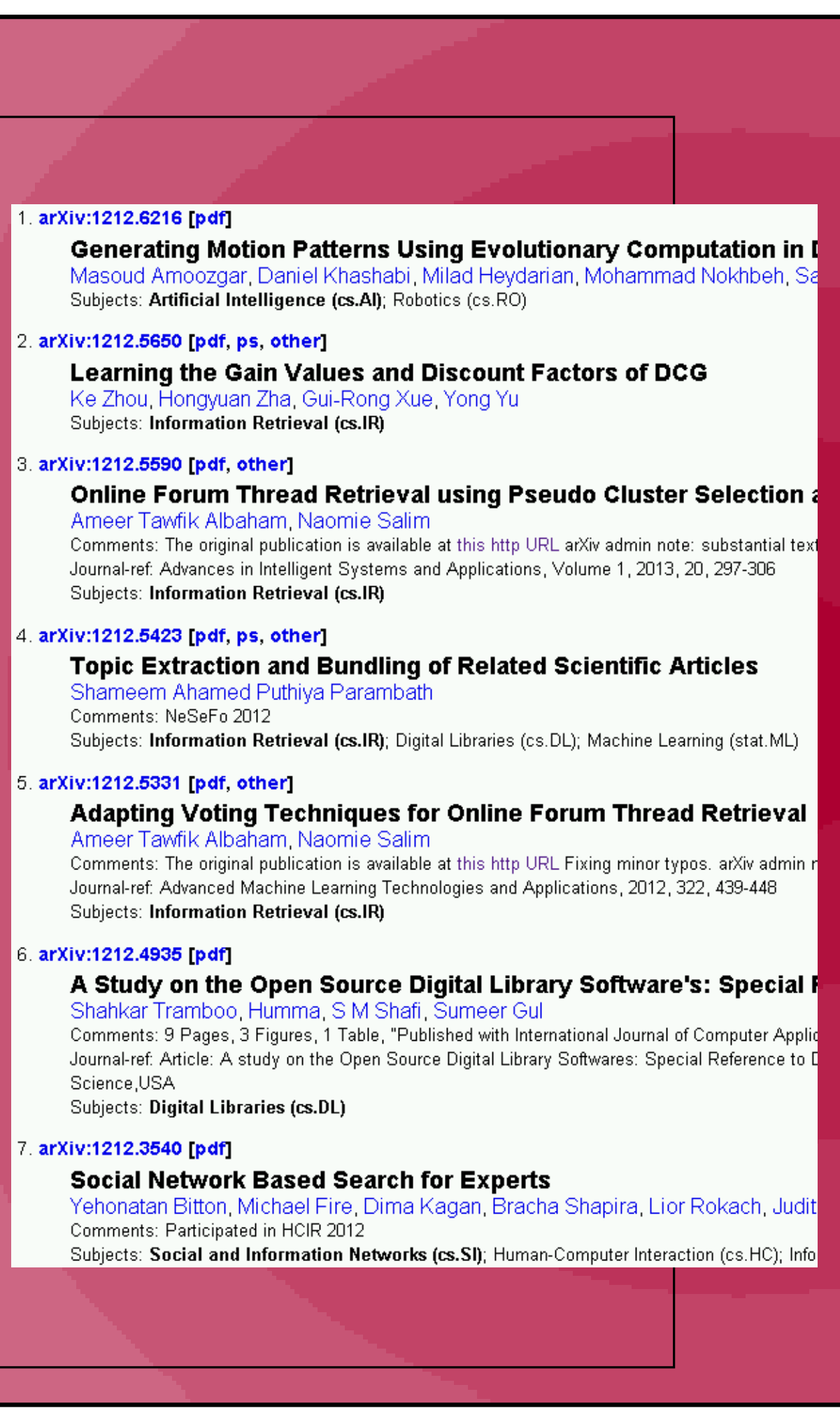
FedWeb 2012

108 resources

Top 10 results
(snippets + pages)

Dec 2011 – Jan 2012

Queries


1. arXiv:1212.6216 [pdf]
Generating Motion Patterns Using Evolutionary Computation in D
Masoud Amoozgar, Daniel Khashabi, Milad Heydarian, Mohammad Nokhbeh, Sa
Subjects: Artificial Intelligence (cs.AI); Robotics (cs.RO)

2. arXiv:1212.5650 [pdf, ps, other]
Learning the Gain Values and Discount Factors of DCG
Ke Zhou, Hongyuan Zha, Gui-Rong Xue, Yong Yu
Subjects: Information Retrieval (cs.IR)

3. arXiv:1212.5590 [pdf, other]
Online Forum Thread Retrieval using Pseudo Cluster Selection a
Ameer Tawfik Albaham, Naomie Salim
Comments: The original publication is available at this http URL arXiv admin note: substantial text
Journal-ref: Advances in Intelligent Systems and Applications, Volume 1, 2013, 20, 297-306
Subjects: Information Retrieval (cs.IR)

4. arXiv:1212.5423 [pdf, ps, other]
Topic Extraction and Bundling of Related Scientific Articles
Shameem Ahamed Puthiya Parambath
Comments: NeSeFo 2012
Subjects: Information Retrieval (cs.IR); Digital Libraries (cs.DL); Machine Learning (stat.ML)

5. arXiv:1212.5331 [pdf, other]
Adapting Voting Techniques for Online Forum Thread Retrieval
Ameer Tawfik Albaham, Naomie Salim
Comments: The original publication is available at this http URL Fixing minor typos. arXiv admin r
Journal-ref: Advanced Machine Learning Technologies and Applications, 2012, 322, 439-448
Subjects: Information Retrieval (cs.IR)

6. arXiv:1212.4935 [pdf]
A Study on the Open Source Digital Library Software's: Special R
Shahkar Tramboo, Humma, S M Shafi, Sumeer Gul
Comments: 9 Pages, 3 Figures, 1 Table, "Published with International Journal of Computer Applic
Journal-ref: Article: A study on the Open Source Digital Library Softwares: Special Reference to D
Science,USA
Subjects: Digital Libraries (cs.DL)

7. arXiv:1212.3540 [pdf]
Social Network Based Search for Experts
Yehonatan Bitton, Michael Fire, Dima Kagan, Bracha Shapira, Lior Rokach, Judit
Comments: Participated in HCIR 2012
Subjects: Social and Information Networks (cs.SI); Human-Computer Interaction (cs.HC); Info

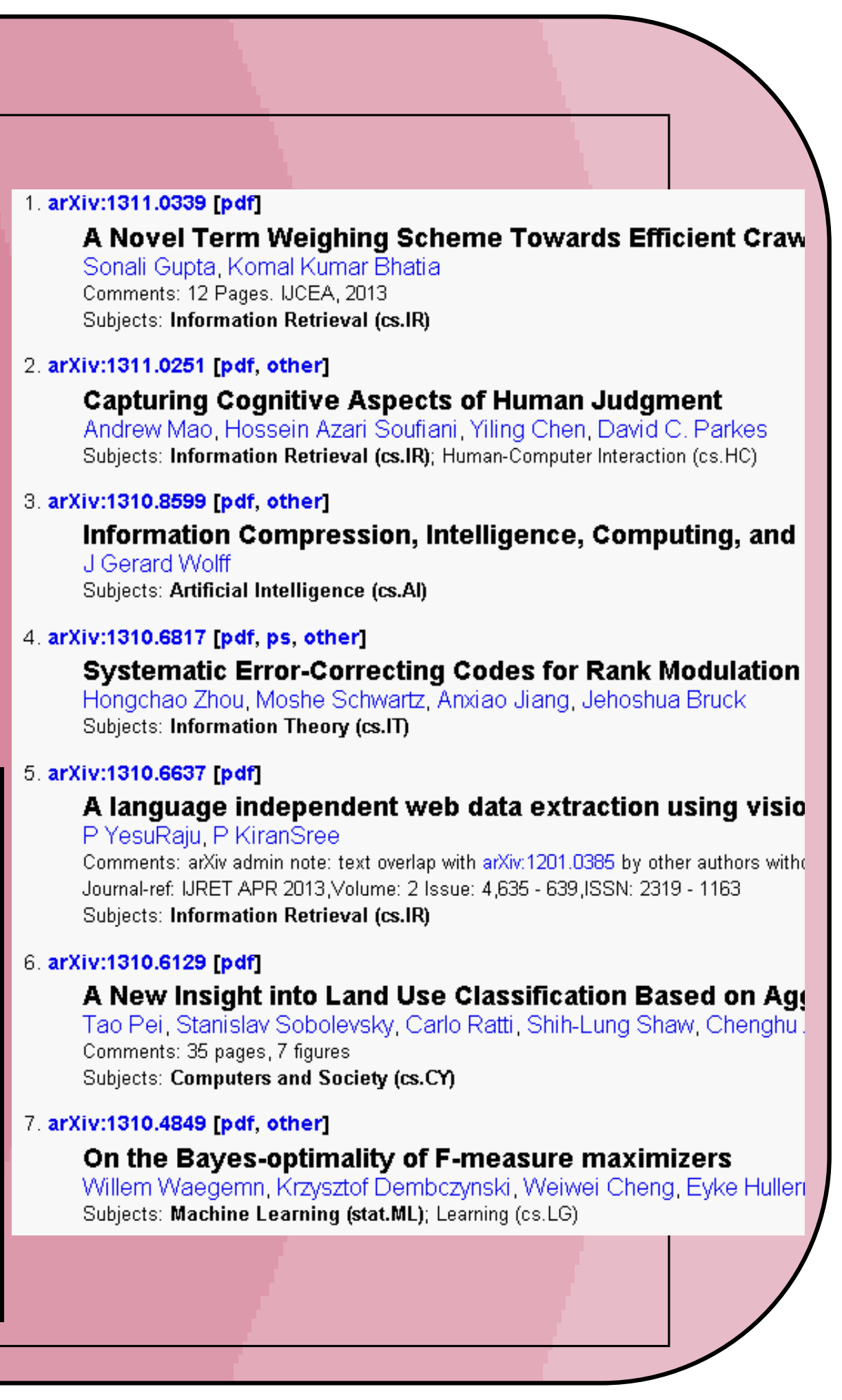
FedWeb 2013

156 + 1 resources
(specific + BigWeb)

Top 10 results
(snippets + pages)

Apr – May 2013

Queries


1. arXiv:1311.0339 [pdf]
A Novel Term Weighing Scheme Towards Efficient Craw
Sonali Gupta, Komal Kumar Bhatia
Comments: 12 Pages. IJCEA, 2013
Subjects: Information Retrieval (cs.IR)

2. arXiv:1311.0251 [pdf, other]
Capturing Cognitive Aspects of Human Judgment
Andrew Mao, Hossein Azari Soufiani, Yiling Chen, David C. Parkes
Subjects: Information Retrieval (cs.IR); Human-Computer Interaction (cs.HC)

3. arXiv:1310.8599 [pdf, other]
Information Compression, Intelligence, Computing, and
J Gerard Wolff
Subjects: Artificial Intelligence (cs.AI)

4. arXiv:1310.6817 [pdf, ps, other]
Systematic Error-Correcting Codes for Rank Modulation
Hongchao Zhou, Moshe Schwartz, Anxiao Jiang, Jehoshua Bruck
Subjects: Information Theory (cs.IT)

5. arXiv:1310.6637 [pdf]
A language independent web data extraction using visio
P YesuRaju, P KiranSree
Comments: arXiv admin note: text overlap with arXiv:1201.0385 by other authors witho
Journal-ref: IJRET APR 2013,Volume: 2 Issue: 4,635 - 639,ISSN: 2319 - 1163
Subjects: Information Retrieval (cs.IR)

6. arXiv:1310.6129 [pdf]
A New Insight into Land Use Classification Based on Agg
Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Chenghu
Comments: 35 pages, 7 figures
Subjects: Computers and Society (cs.CY)

7. arXiv:1310.4849 [pdf, other]
On the Bayes-optimality of F-measure maximizers
Willem Waegemn, Krzysztof Dembczynski, Weiwei Cheng, Eyke Hullern
Subjects: Machine Learning (stat.ML); Learning (cs.LG)

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

Results merging

Relevance

Documents ranked with respect to the query likelihood model:

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

Run: IndriQL

Cluster

1. Rank resources (previous task)
2. Documents within a resource are ranked with IndriQL

Run: clTODPJ. Not submitted

Diversity

IA-select diversification of IndriQL ranking using query relevance with respect to the resources

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

resource

Run: iaTODPJ

Boost

Use directly the relevance with respect to the resources to boost the documents

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

Run: bstTODPJ

Discussion

Results merging can be solved with simple IR techniques

Query likelihood obtains very good results

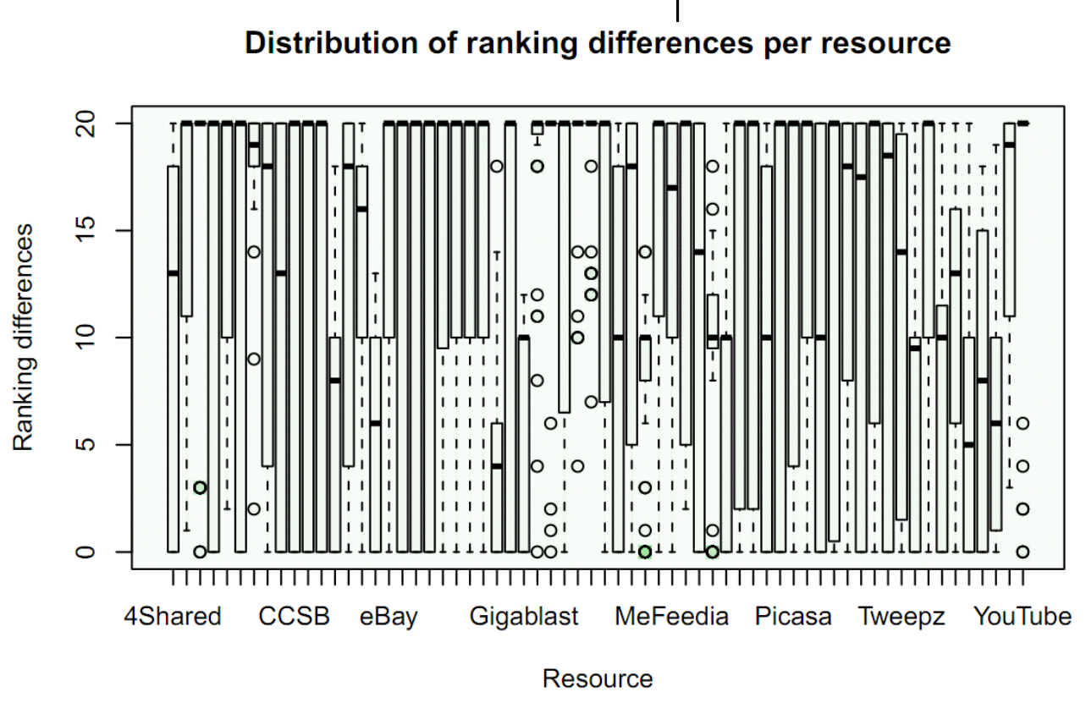
How to define a training set for an evolving test environment?

The rankings of the resources change

The content of the websites change

The type of queries is important:

are they tailored to be answered by a specific resource?



Text REtrieval Conference (TREC 2013)
Gaithersburg, Maryland, USA, 19-22 November 2013