CWI @ TREC 2013: Contextual Suggestion Track

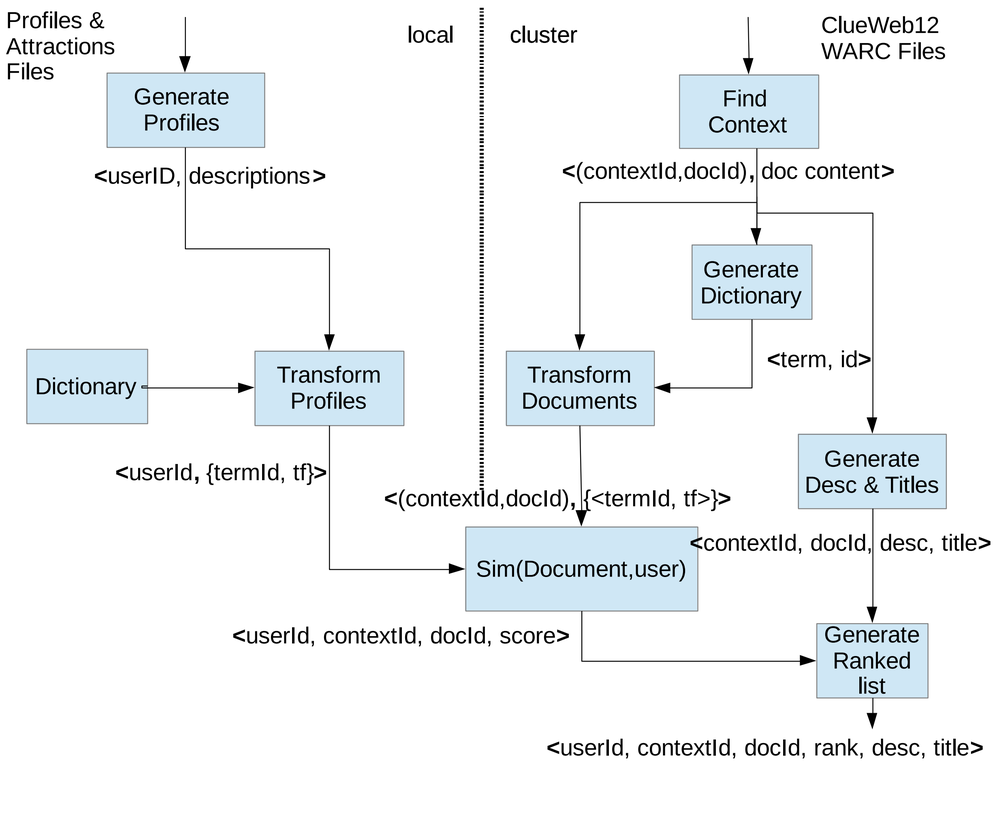
Thaer Samar†, Alejandro Bellogín†, Arjen P. de Vries†, Jimmy Lin‡, Alan Said†

† Centrum Wiskunde & Informatica, The Netherlands

‡ University of Maryland, United States

{t.m.h.samar, a.bellogin, arjen.de.vries, alan.said}@cwi.nl, jimmylin@umd.edu

General framework



Input

ClueWeb 12

5.5 / 27.3 TB

33,447 WARC files

733,019,372 documents

Profiles and attractions

562 user profiles

50 attractions in Philadelphia, PA

28,100 ratings to description and site

Filtering

Exact mention of the context

Format: <City, ST>

Exclude

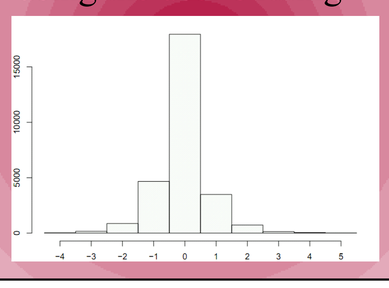
non related sentences

mention the city in other state

Found 13,548,982 documents

Profile generation

Based on attraction's description using the user's rating



Representation and transformation

Vector space model

Elements of the vectors are <term, frequency> pairs

Efficient in terms of

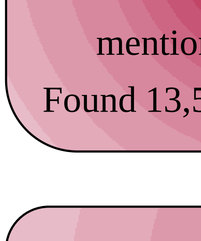
Size: from 918 GB to 40 GB

Processing speed

document vector

user vector

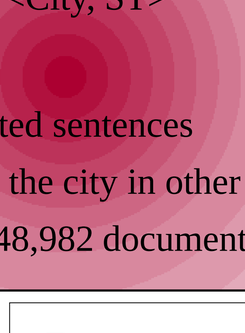
cosine similarity



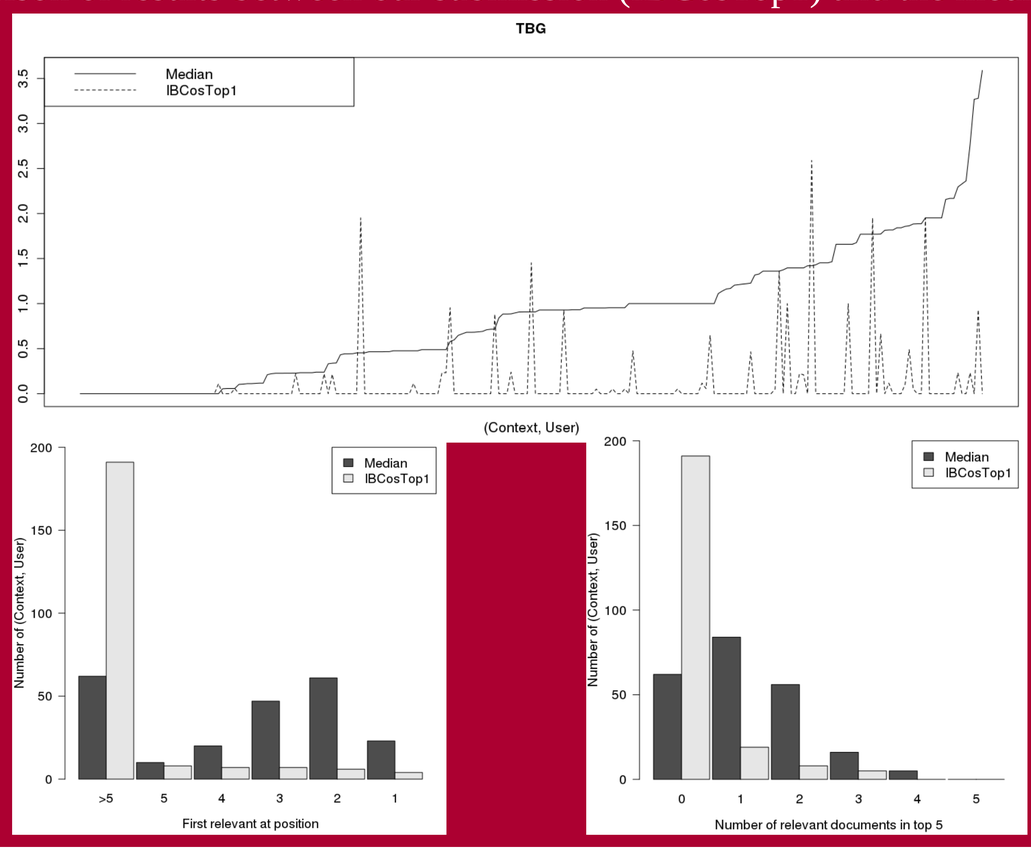
museum

wellness

music



Comparison of results between our submission (IBCosTop1) and the median



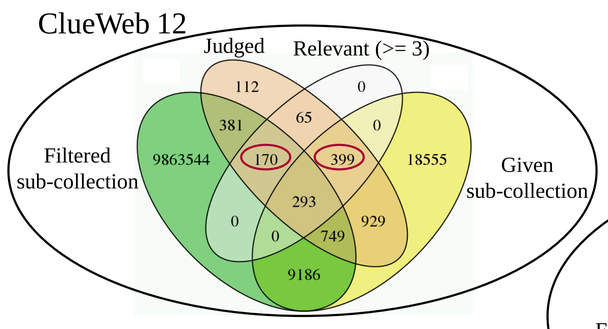
Discussion

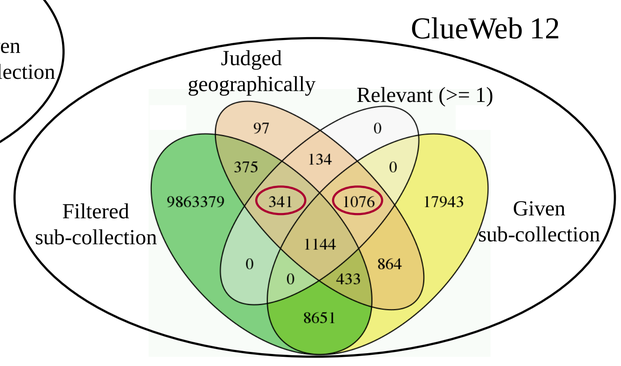
Effect of filtering

Geographically appropriate documents are missing

Method	MRR	MRR_d	$P@5_d$	$P@5_{d\bar{g}}$
IBCosTop1	**0.0559**	0.0745	**0.0587**	**0.2202**
IBCosTop1 (given)	0.0528	**0.0955**	0.0484	0.0780

A sub-collection provided by the organisers (*given*) has a higher coverage of geographically appropriate documents than our filtered sub-collection





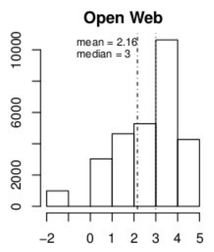
Effect of personalisation

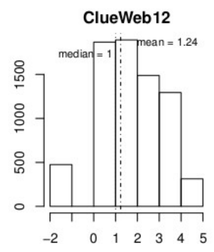
Not easy to compare: low coverage of relevance assessments

Method	MRR	MRR_d	$P@5_d$	$P@5_{d\bar{g}}$
IBCosTop1	**0.0559**	**0.0745**	**0.0587**	**0.2202**
IBCosTop1 + 5NN text cos	0.0455	0.0562	0.0330	0.1486
IBCosTop1 + 5NN text Jacc	0.0433	0.0521	0.0330	0.1294
IBCosTop1 + 5NN rating cos	0.0429	0.0553	0.0349	0.1477
IBCosTop1 + 5NN rating Pearson	0.0450	0.0580	0.0358	0.1560
Classifier + 5NN text cos	0.0045	0.0112	0.0036	0.0251
Classifier + 5NN text Jacc	0.0045	0.0121	0.0045	0.0260
Classifier + 5NN rating cos	0.0045	0.0090	0.0027	0.0242
Classifier + 5NN rating Pearson	0.0045	0.0067	0.0018	0.0233
Positive profile	0.0396	0.0588	0.0359	0.1498
Negative profile	0.0045	0.0045	0.0009	0.0152
Positive + 5NN text cos	0.0426	0.0572	0.0341	0.1399

Effect of evaluation

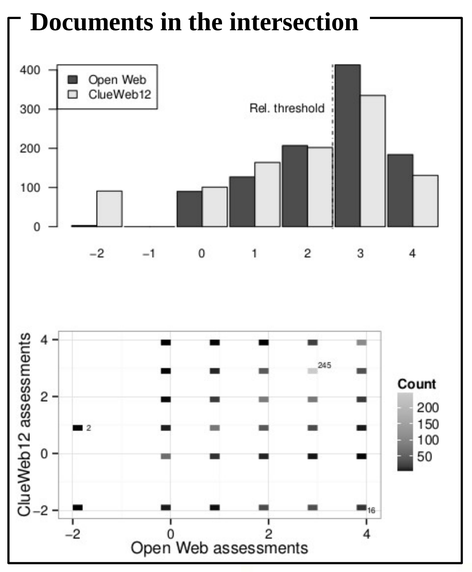
We found a bias towards methods using documents from the Open Web





Collection	Method	P@5	MRR	TBG
Open Web	Oracle + geo	**0.909**	**0.945**	**4.030**
Open Web	Oracle	0.742	0.845	2.767
ClueWeb12	Oracle + geo	0.509	0.761	2.221
ClueWeb12	Oracle	0.413	0.640	1.422
ClueWeb12 sub	Oracle + geo	0.418	0.702	1.870
ClueWeb12 sub	Oracle	0.393	0.652	1.566

Documents in the intersection



Text REtrieval Conference (TREC 2013)

Gaithersburg, Maryland, USA, 19-22 November 2013

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