Challenges on Combining Open Web and Dataset Evaluation Results: The Case of the Contextual Suggestion Track

Alejandro Bellogín[†,‡], Thaer Samar[†], Arjen P. de Vries[†], Alan Said[†]

[†] Centrum Wiskunde & Informatica, The Netherlands

[‡] Universidad Autónoma de Madrid, Spain

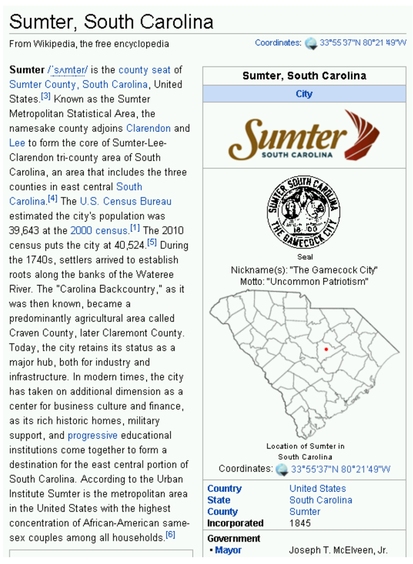
alejandro.bellogin@uam.es, {samar, arjen, alan}@cwi.nl

Contextual Suggestion

museum

wellness

music



Input

Profiles and attractions

562 user profiles

50 attractions in Philadelphia, PA

28,100 ratings to description and site

Suggestions coming from

Open Web

Any URL available

Typically: Yelp, Google Places, Foursquare, TripAdvisor

ClueWeb12 dataset

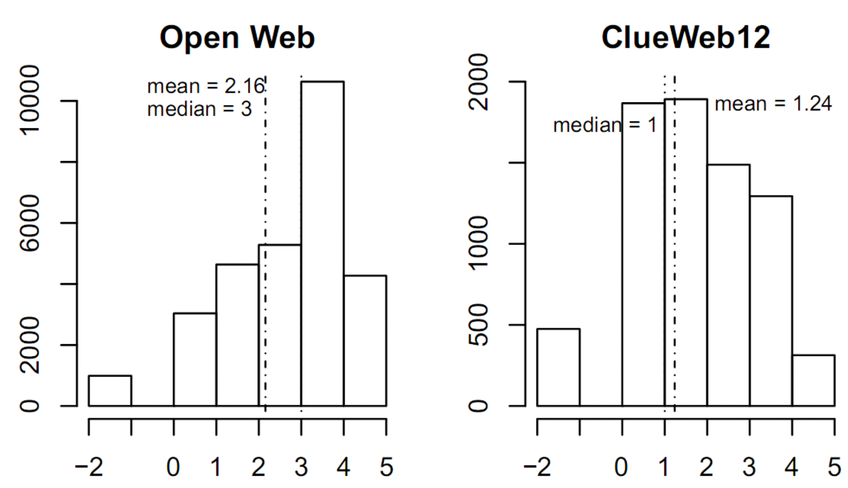
> 700 million Web pages

Between Feb – May 2012

Challenges

Fair comparison of datasets

Judgements for documents from the Open Web are skewed towards the positive (relevant) values.



Even though ClueWeb12 was seeded with travel sites, Yelp (which provided the highest number of relevant documents) is not included due to very strict crawling rules

Dynamic websites, pages within social networks, and fresh content (after ClueWeb12 was collected) are among the documents retrieved using the Open Web and not contained in ClueWeb12

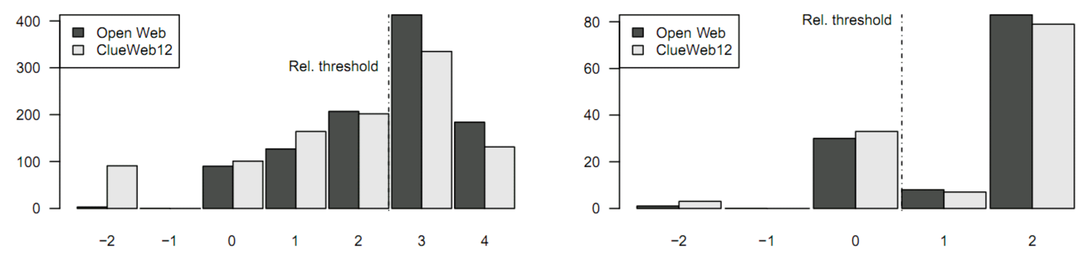
An oracle method based exclusively on documents from the Open Web achieves higher performance.

Collection	Method	P@5	MRR	TBG	$P@5_r$	MRR_r
Open Web	Oracle + geo	**0.909**	**0.945**	**4.030**	**0.950**	0.957
Open Web	Oracle	0.742	0.845	2.767	**0.950**	**0.962**
ClueWeb12	Oracle + geo	0.509	0.761	2.221	0.700	0.892
ClueWeb12	Oracle	0.413	0.640	1.422	0.700	0.892
ClueWeb12 sub	Oracle + geo	0.418	0.702	1.870	0.551	0.803
ClueWeb12 sub	Oracle	0.393	0.652	1.566	0.557	0.814

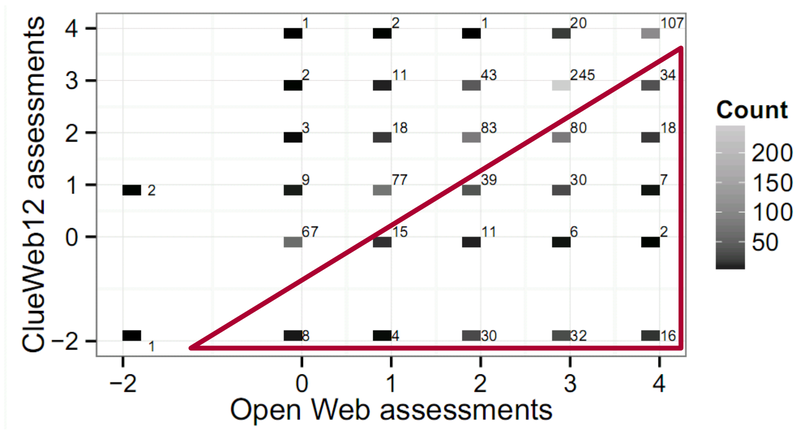
This result is consistent when no geographical information is used to account for relevance ($P@5_r$ and MRR_r) and when a specifically tailored ClueWeb12 subcollection (ClueWeb12 sub) is tested.

Consistency of evaluation judgements

The subset of documents of ClueWeb 12 also submitted as Open Web documents tend to differ in terms of subjective assessments (profile relevance, left), whereas they receive consistent objective assessments (geographical relevance, right).



The frequency of two inconsistent assessments is particularly detrimental for ClueWeb12 documents (inside the triangle).



Part of these differences in perceived quality can be attributed to a different rendering of the documents in each dataset.

Conclusion

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

A static, archived dataset is prone to be in disadvantage with respect to the dynamic, live Open Web dataset

The assessor should not be aware of the origin of the document being evaluated

Thanks to

WebART Web Archive Retrieval Tools

ERCIM The European Research Consortium for Informatics and Mathematics

COMMIT/

36th European Conference on Information Retrieval (ECIR 2014)

Amsterdam, The Netherlands, 13-16 April 2014