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

Alejandro Bellogín[1,2], Thaer Samar[1], Arjen P. de Vries[1], and Alan Said[1]

[1] Centrum Wiskunde & Informatica, {samar,arjen,alan}@cwi.nl
[2] Universidad Autónoma de Madrid, alejandro.bellogin@uam.es

Abstract. The TREC 2013 Contextual Suggestion Track allowed participants to submit personalised rankings using documents either from the Open Web or from an archived, static Web collection, the ClueWeb12 dataset. We argue that this setting poses problems in how the performance of the participants should be compared. We analyse biases found in the process, both objective and subjective, and discuss these issues in the general framework of evaluating personalised Information Retrieval using dynamic against static datasets.

1 Introduction and Motivation

The Contextual Suggestion TREC Track investigates search techniques for complex information needs that are highly dependent on context and user interests. Input to the task are a set of profiles (*users*), a set of example suggestions (*attractions*), and a set of contexts (*locations*). Each attraction includes a title, a description, and an associated URL. Each profile corresponds to a single user, and indicates the user's preference with respect to each attraction. Two ratings are used: one for the attraction's title and description and another one for its website. Finally, each context corresponds to a particular geographical location (a city and its corresponding state in the United States). With this information, up to 50 ranked suggestions should be generated by each participant for every context and profile pair. Each suggestion should be appropriate to both the user's profile and the context. The description and title of the suggestion may be tailored to reflect the preferences of that user.

As opposed to the 2012 track, where only submissions from the Open Web were allowed [2], in 2013, the organisers introduced the option of submitting rankings from one of the TREC collections. In particular, from the ClueWeb12 dataset, which includes more than 700 million Web pages crawled from the Internet between February and May 2012. This approach would allow further comparison and reproducibility of the results, in agreement with most of the other TREC tracks developed so far.

The Contextual Suggestion track has, however, some characteristics that challenge the standard TREC evaluation setting, and the Cranfield paradigm at large. First, relevance is defined as a bi-dimensional variable, since a document has to be interesting for the user and appropriate for the given context. Second, it is personalised, hence the typical pooling mechanism where several

judges are used and their judgements are aggregated cannot be used. Third, and as a consequence of the second, there is no explicit query in the system, or, as stated in [1], the query "entertain me" is implicitly assumed and fixed during the retrieval/recommendation stage.

In this paper we analyse whether the evaluation results and relevance assessments obtained in this year's track (2013) may be hiding the fact that the ClueWeb12 dataset does not contain as many *interesting* documents as an Open Web dataset. We show that there is an (implicit) bias to receive better judgements by the Open Web submissions when compared with those using ClueWeb12. This result is evidenced by comparing a subset of documents which was assessed as Open Web and ClueWeb12 documents separately, ending up with inconsistent assessments. We finally discuss more general connotations of this work in the broader context of evaluating interactive Information Retrieval with users, combining documents from live and archived web.

2 Experimental Setup

In our analysis, we use ground truth relevance assessments provided by the organisers of the TREC 2013 Contextual Suggestion track. These are provided as two separate categories, depending on whether the relevance is personal (how interesting is this document for the user, in a particular context) or related to the geographical appropriateness of the document in the given context. These judgements also have different scales: subjective judgements range from 0 (strongly uninterested) to 4 (strongly interested) whereas objective judgements go from 0 (not geographically appropriate) to 2 (geographically appropriate). Besides, in both cases, a value of -2 indicates that the document did not load.

Out of the $28,100$ possible combinations of user profiles (562) and contexts (50), only 223 (user, context) pairs were evaluated, roughly 0.8% of the complete set of pairs. For these pairs, the top-5 documents of every submission were judged by the users (profile and geographical relevance) and NIST assessors (geographical relevance). The metrics used to evaluate the submissions were Precision at 5 (P@5), Mean Reciprocal Rank (MRR), and Time-Biased Gain (TBG) [1]. These metrics consider the geographical and profile relevance (both in terms of document and description judgement), taking as thresholds a value of 1 and 3 (inclusive), respectively.

3 Challenges

In this section we present the analysis derived from the data described above. We first compare Open Web and ClueWeb12 in general, to see whether one of them has a higher inherent quality than the other (fair comparison of datasets). Next, we performed a pairwise comparison in a subset of documents shared by the two datasets (consistency of evaluation judgements).

3.1 Fair Comparison of Datasets

We start by analysing how comparable – in terms of inherent quality – the Open Web and ClueWeb12 datasets are for the Contextual Suggestion task. This is

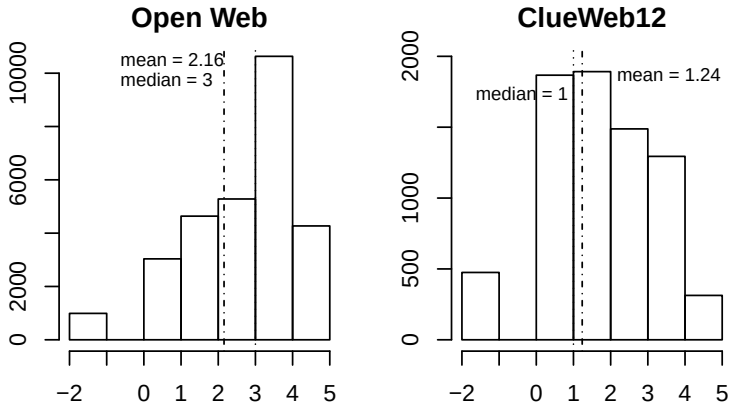

Fig. 1. Histogram of document judgements for documents from the Open Web (left) and from ClueWeb12 (right).

important because if, for some reason, one of the datasets is richer, or has a biased view of the world, the documents retrieved will share such a bias. This is, in principle, not desirable [4].

For this analysis, we leave out the user, context, and system variables, and compare the judgements given to documents from the Open Web against those from ClueWeb12. We can see the results in Fig. 1. We observe that the Open Web histogram is slightly skewed towards the positive, relevant judgements. Even though we are not interested in comparing the actual frequencies (this would not be fair, mainly because there were many more Open Web submissions than ClueWeb12 ones[3]), it is still relevant to see the relative frequency of -2's and 5's in each dataset: this is an important difference which will impact the performance of the systems using ClueWeb12 documents, as we shall see next.

This first check may indicate that, although the ClueWeb12 dataset is a snapshot of the Web, it does not contain as many *entertaining* or interesting documents as those datasets specially tailored from the Open Web for this task. In fact, such datasets are typically seeded from candidate sources known for having interesting content, such as Yelp, Google Places, Foursquare, and TripAdvisor [2], in contrast to ClueWeb12, which aims to cover a more uniform set of topics; even though it was seeded with specific travel sites. This apparent contradiction is explained by the fact that ClueWeb12 is missing Yelp due to very strict indexing rules (see http://yelp.com/robots.txt). It turns out that this site alone provided the highest number of relevant documents in the two editions of the track. In this situation, the fact that one of the datasets lacks a specific bias which affects the other dataset and, likely, the ultimate user need – "entertain me" – seems to limit its applicability in this context.

As an additional test, we build *oracle* submissions based on the relevance judgements (both subjective and objective). Although we have observed that Open Web documents tend to receive higher ratings, this does not necessarily mean that the *best possible performance* of the different methods should be very different. Table 1 shows the performance values obtained by evaluating

[3] More specifically, 27 runs submitted URLs from the Open Web, and only 7 used ClueWeb12 documents.

Table 1. Comparison of performance for different methods based on the data used to derive the best ranking. The metrics P@5$_r$ and MRR$_r$ denote that no geographical information is used to account for relevance.

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

the rankings found when the documents are ordered according to subjective and objective assessments (*Oracle + geo*) or only the subjective judgements (*Oracle*). We also present the metrics P@5$_r$ and MRR$_r$ where the geographical (objective) information is not considered. We generated the rankings using three subsets of the assessed documents: those submitted as Open Web, those submitted as ClueWeb12, and among those submitted as ClueWeb12, those coming from a subcollection (ClueWeb12 sub) provided by the organisers. Since each of these methods may potentially generate a different description, in the evaluation of these rankings we do not consider the description judgements (i.e., they are assumed to be always relevant), which should not affect the observed ranking of the methods since the same setting is applied for all of them.

Based on the results presented in Table 1, we find that the performance of rankings using exclusively documents from the Open Web is always the highest. Specifically, the drop in performance seems to be quite significant; for instance, in terms of P@5, the lowest value for Open Web is 0.742, whereas the highest value using ClueWeb12 is 0.509. Furthermore, even when a subcollection from ClueWeb12 specifically tailored for the task[4] is used, the results do not improve at all.

As we discussed before, a possible reason for these values is that ClueWeb12 does not contain as many interesting documents as a tailored subset of the Web. Specifically, we found less than 20% of the assessed Open Web documents appearing in ClueWeb12. Dynamic websites, pages within social networks (e.g., Google+, Facebook), or fresh content (created after the ClueWeb12 was crawled) mostly correspond to these missed documents.

3.2 Consistency of Evaluation Judgements

While doing the aforementioned analysis, we identified a subset of documents that were submitted as part of the ClueWeb12 dataset whose corresponding URLs were also submitted (by other participants) as Open Web documents. Since both identifiers (URL and ClueWeb12 id) correspond to the same document/webpage, we decided to investigate if we could detect any bias towards one of the datasets based on this sample of the judgements. Hence, out of the 36,178 available assessments, we ended up with 1,024 corresponding to the same

[4] As described by the organisers in https://sites.google.com/site/treccontext/trec-2013/subcollection: "this subcollection was created by issuing a variety of queries targeted to the Contextual Suggestion track on a commercial search engine."

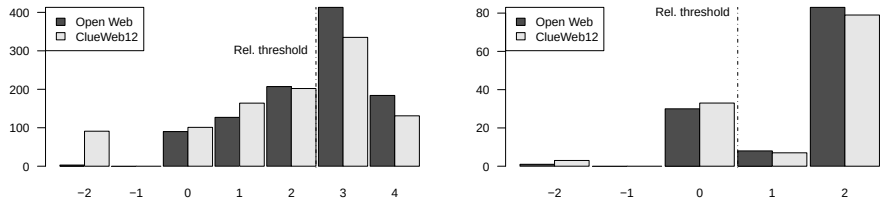

Fig. 2. Histogram of document judgements for the same documents from the Open Web and from ClueWeb12. In the left figure, the ratings show subjective relevance (interest), whereas the right one shows the objective relevance of the document (geographically appropriate).

document retrieved by different systems for exactly the same user and context, and thus, the relevance judgement should be, in principle, the same.

Fig. 2 shows the histograms of the subjective (left) and objective (right) assessments on the sampled documents. We notice how the shape of the objective assessments (geographical relevance) is very similar in both datasets, whereas that of the subjective assessments (profile relevance) has some differences, especially in the number of 3's, 4's, and -2's. To further emphasise this difference, we performed a Wilcoxon signed-rank significance test where the null hypothesis is that both variables have the same mean. We obtained a p-value of 0.5018 for the first case, and a value close to 0 for the second case. This indicates that we cannot reject the null hypothesis in the former, whereas in the latter it is very unlikely that the null hypothesis is true. This justifies a conclusion that the subjective assessments from Open Web and ClueWeb12 are very likely to be different (while this problem does not exist for the geographical relevance case). To provide some context, the consistency of these assessments within each dataset (Open Web and ClueWeb12) as measured by the standard deviation is around 0.05.

An alternative visualisation of this situation is depicted in Fig. 3, where we show in a scatterplot the judgements received by the same document in each dataset. In this figure we can also observe how frequent these combinations occur. For instance, in 245 cases, a document assessed as 3 in the Open Web has also been assessed as 3 in ClueWeb12. It is important to note that ClueWeb12 documents receive a larger amount of -2 judgements (as we could see in Fig. 2 left) almost independently of the corresponding assessment received by the Open Web document. Hence, in the extreme situation, 16 times an Open Web document assessed as 4 received a -2 as a ClueWeb12 document. This behaviour, however, is neglible for low assessed Open Web documents in the inversed situation.

Part of the differences in judgements can be attributed to a different rendering of the document for each dataset[5]. Assessors are influenced by several conditions, one of them is the visual aspect of the interface, but also the response time, the order of examination, the familiarity with the interface, etc. [3]. Therefore, it is important that these details are kept as stable as possible when different datasets are evaluated at the same time.

[5] Confirmed via email with the organisers.

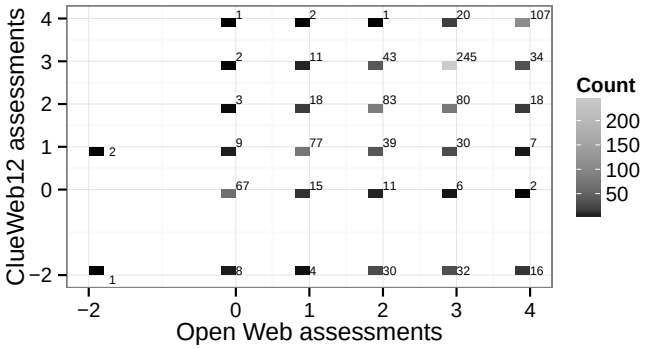

Fig. 3. Scatterplot showing the frequencies of the available assessments combinations for the same document (by the same user in the same context), being judged as a document from the Open Web or from ClueWeb12.

4 Conclusions

We analysed two concrete challenges for measuring effectiveness of information retrieval systems on a dataset where the documents originate from a mix of sources. A fair comparison would require a similar ratio of potential candidate documents to be relevant. Additionally, when the considered user need is time-sensitive (as in our case), a static, archived dataset is prone to be in disadvantage with respect to a dynamic, live Open Web dataset.

Consistency in the judgements is also an important challenge, not always possible due to incomplete mappings between the datasets at hand. Special efforts are therefore required to design a user experience where the assessor is not aware of the origin of the document being evaluated. Aspects such as timing and fatigue should also be considered [3].

Acknowledgments. Part of this work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme, funded by European Comission FP7 grant agreement no.246016. This research was further supported by the Netherlands Organization for Scientific Research (NWO project 640.005.001) and the Dutch national program `COMMIT/`.

References

1. Dean-Hall, A., Clarke, C.L.A., Kamps, J., Thomas, P.: Evaluating contextual suggestion. In: Proceedings of the Fifth International Workshop on Evaluating Information Access (EVIA 2013) (2013)
2. Dean-Hall, A., Clarke, C.L.A., Kamps, J., Thomas, P., Voorhees, E.: Overview of the TREC 2012 contextual suggestion track. In: Proceedings of the Twenty First Text REtrieval Conference (TREC 2012). NIST (2013)
3. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval 3(1-2), 1–224 (2009)
4. Minka, T., Robertson, S.: Selection bias in the LETOR datasets. In: Proceedings of SIGIR 2008 Workshop on Learning to Rank for Information Retrieval (2008)