Clarity Re-Visited

Shay Hummel, Anna Shtok,
Fiana Raiber, Oren Kurland
Faculty of Industrial Engineering and
Management, Technion, Haifa 32000, Israel
{hummels,annabel,fiana}@tx.technion.ac.il,
kurland@ie.technion.ac.il

David Carmel
IBM Research Lab, Haifa 31905, Israel
carmel@il.ibm.com

ABSTRACT

We present a novel interpretation of **Clarity** [5], a widely used query performance predictor. While Clarity is commonly described as a measure of the "distance" between the language model of the top-retrieved documents and that of the collection, we show that it actually quantifies an additional property of the result list, namely, its diversity. This analysis, along with empirical evaluation, helps to explain the low prediction quality of Clarity for large-scale Web collections.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Algorithms, Experimentation

Keywords: query-performance prediction, Clarity

1. INTRODUCTION

Many query performance predictors were proposed over the years [2]. Clarity [5] is a well known, commonly used, state-of-the-art predictor that measures the "coherence" of the top-retrieved documents with respect to the collection. Specifically, the more distinguishable the language used in the retrieved documents from the general language used in the collection, the better the retrieval is assumed to be[1]. Clarity was shown to be highly effective for most TREC benchmarks [6]. However, low prediction quality is observed when using Clarity for large scale, noisy, Web corpora [1].

We present a novel formal analysis of Clarity that sheds some light on its underlying components and the properties of the result list of top-retrieved documents that it quantifies. While Clarity is commonly described as a measure of the "distance" between a language model induced from the result list and that induced from the collection, we show that Clarity actually quantifies an additional property of the result list, namely, its diversity. Our empirical analysis shows that the diversity of the result list has a negative correlation with retrieval performance for older TREC benchmarks and a positive correlation for the new ClueWeb collection. These findings, along with the formal analysis, help to explain the poor prediction quality of Clarity over ClueWeb. In addition, the novel interpretation we present suggests new integration approaches of Clarity's building blocks.

[1] There are are several variants of clarity, among which is a pre-retrieval method (SCS [7]) that considers only the query and the corpus and not the retrieved documents.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. INSIDE CLARITY

Let q and $\mathcal{D}$ denote a query and a corpus of documents, respectively. In what follows we use $p(w|x)$ to denote the probability assigned to term w by a unigram (smoothed) language model induced from x.

The query likelihood (QL) retrieval method scores document d by $\log p(q|d) \overset{def}{=} \log \prod_{q_i \in q} p(q_i|d)$, where q_i is a term in q. Let $\mathcal{D}_q^{[k]}$ denote the result list of the k highest ranked documents. The assumption behind Clarity is that the higher the divergence of a model of $\mathcal{D}_q^{[k]}$ from that of the corpus, the higher the effectiveness of $\mathcal{D}_q^{[k]}$ is, and thereby, the better the quality of the QL-based retrieval. The KL divergence between a relevance language model, R, induced from $\mathcal{D}_q^{[k]}$, and a language model induced from $\mathcal{D}$, is used to quantify this divergence. R is a weighted linear mixture of the models of documents in $\mathcal{D}_q^{[k]}$ [5]. Accordingly, the Clarity of q is defined as:

$$Clarity(q) \overset{def}{=} KL\left(p(\cdot|R) \;\middle|\middle|\; p(\cdot|\mathcal{D})\right) = \sum_w p(w|R) \log \frac{p(w|R)}{p(w|\mathcal{D})}.$$

As is the case for any two probability distributions, we can write the KL divergence as follows:

$$KL\left(p(\cdot|R) \;\middle|\middle|\; p(\cdot|\mathcal{D})\right) = CE\left(p(\cdot|R) \;\middle|\middle|\; p(\cdot|\mathcal{D})\right) - H\left(p(\cdot|R)\right);$$

CE is the cross entropy between R and $\mathcal{D}$,

$$CE\left(p(\cdot|R) \;\middle|\middle|\; p(\cdot|\mathcal{D})\right) \overset{def}{=} -\sum_w p(w|R) \log p(w|\mathcal{D});$$

H is the entropy of the relevance model,

$$H\left(p(\cdot|R)\right) \overset{def}{=} -\sum_w p(w|R) \log p(w|R).$$

Under this decomposition, Clarity integrates two measures (building blocks). The first is the "distance" of R from the corpus, as measured by the cross entropy. The more distant R from $\mathcal{D}$, the higher the cross entropy is. We use **CDistance** (for "corpus distance") to refer to the cross entropy between R and $\mathcal{D}$.

The second measure used by Clarity is the entropy of R. High entropy means that R assigns relatively low weights (i.e., probabilities) to a large number of terms; thereby, $\mathcal{D}_q^{[k]}$ is highly diverse. In contrast, low entropy means that only a few terms are highly weighted, hence $\mathcal{D}_q^{[k]}$ is more focused. We use **LDiversity** (for "list diversity") to refer to R's entropy. Next we study the prediction quality of each of these building blocks and compare it with that of Clarity which amounts to their equal-weight linear interpolation:

$$Clarity(q) = CDistance(q) - LDiversity(q). \qquad (1)$$

3. EXPERIMENTS

We conducted experiments using the following TREC benchmarks (disks and topics are indicated in the parentheses): TREC4 (disks 2-3; 201-250), TREC5 (disks 2,4; 251-300), WT10G (WT10G; 451-550), Robust (disks 4-5 - {CR}; 301-450, 601-700), GOV2 (GOV2; 701-850), and the ClueWeb collection (category B). Two sets of topics were used for ClueWeb: Clue09 (1-50) and Clue10 (51-100). We applied Porter stemming and stopword removal upon queries and documents using the Lemur/Indri toolkit.

Previous work hypothesized that the low prediction quality of Clarity over Web collections is due to the large amount of noise (e.g. spam) [6]. To address spam effects in ClueWeb, we filtered out the spammiest documents from the result lists (those assigned a spam score below 50 by Waterloo's classifier [4]) and retained the original ranking for the residual corpus. Thus, we get two additional experimental setups for ClueWeb: Clue09+SpRM and Clue10+SpRM.

We use the predictors to predict the performance of the QL retrieval method specified above; unigram Dirichlet-smoothed document language models are used with the smoothing parameter set to 1000. We study three predictors: CDistance, LDiversity, and Clarity which interpolates the two (see Equation 1). Following the common practice to evaluating prediction quality [2], we report Pearson's correlation between the values assigned by the predictor and retrieval effectiveness measured by average precision computed using TREC's relevance judgments. The size of the result list, k, was set to 500 following previous recommendations [6]. The relevance models were clipped to use only the 100 highest weighted terms. The language models of documents used to construct the relevance model were not smoothed.

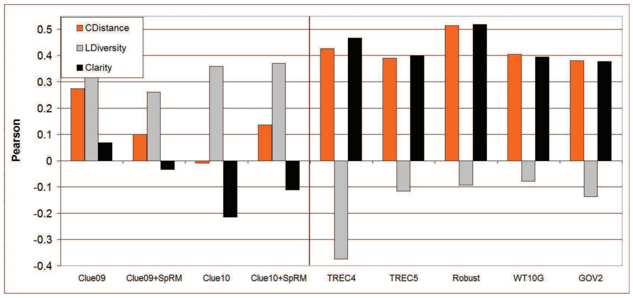

Figure 1: Comparing the prediction quality of Clarity and its building blocks.

Figure 1 presents our main results. To simplify the presentation, we refer to the ClueWeb benchmarks as "ClueWebs" (on the left of the graph) and to all the other benchmarks as "SmallScales" (on the right). The differences of the patterns observed for "ClueWebs" and "SmallScales" are as follows. First, CDistance is not very effective over "ClueWebs" in comparison to "SmallScales". We attribute this finding to the low quality of the collection statistics in a noisy Web setup that affects the estimation of the corpus language model. Evidently, spam removal did not improve prediction quality over ClueWeb.

Second, LDiversity and retrieval performance have positive correlation for "ClueWebs", yet negative correlation is observed for "SmallScales". We presume that list coherence can attest to improved retrieval effectiveness, as is implied by the findings for "SmallScales", which are mainly composed of unambiguous (coherent) queries. On the other hand, list diversity might correspond to improved retrieval performance, as is implied by the findings for "ClueWebs", for ambiguous queries, if it attests to coverage of various query aspects. It was found that for Clue09 topical diversity and retrieval performance are strongly correlated [3].

Following the observations made above, we can explain the low effectiveness of Clarity for "ClueWebs"; Clarity, as presented in Equation 1, is the **subtraction** of prediction values assigned by two predictors which are both positively correlated with retrieval effectiveness. Usually, two predictors that are positively correlated with retrieval performance are integrated by multiplication or summation [2]. Thus, the **subtractive** integration of CDistance and LDiversity, as implemented by Clarity, yields low quality prediction over "ClueWebs".

4. SUMMARY

We showed that Clarity amounts to an equal weight interpolation of two predictors; one measures the "distance" of the result list from the collection, while the second measures the list's "diversity". We used this formal finding to help explain the low prediction quality of Clarity over ClueWeb, in contrast to its high effectiveness over other TREC benchmarks. Preliminary results of using non-equal weights for the interpolation mentioned above, and independently optimizing free-parameter values for each predictor, attest to the merits of these future directions. (Actual numbers are omitted due to space considerations.)

Acknowledgments. We thank the reviewers for their comments. This paper is based upon work supported in part by the Israel Science Foundation under grant no. 557/09, by IBM's SUR award, by an IBM Ph.D. Fellowship and by Miriam and Aaron Gutwirth Memorial Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

5. REFERENCES

[1] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Predicting query performance on the web. In *Proceedings of SIGIR*, pages 785–786, 2010.

[2] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval.* Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool, 2010.

[3] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proceedings of TREC*, 2009.

[4] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of SIGIR*, pages 299–306, 2002.

[6] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proceedings of CIKM*, pages 439–448, 2008.

[7] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proceedings of SPIRE*, pages 43–54, 2004.