Evaluation of Resource Description Quality Measures

Mark Baillie[*]
i-lab: CIS Department
University of Strathclyde
Glasgow, UK

mb@cis.strath.ac.uk

Leif Azzopardi[†]
ILPS Group
University of Amsterdam
Netherlands

leif@science.uva.nl

Fabio Crestani
i-lab: CIS Department
University of Strathclyde
Glasgow, UK

fabioc@cis.strath.ac.uk

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Models*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval —*Retrieval Models*

Keywords

Distributed Information Retrieval, Evaluation Metrics

1. INTRODUCTION

An open problem for Distributed Information Retrieval is how to represent large document repositories (known as resources) efficiently. To facilitate resource selection, estimated descriptions of each resource are required, especially when faced with non-cooperative distributed environments[1]. Accurate and efficient Resource description estimation is required as this can have an affect on resource selection, and as a consequence retrieval quality. Query-Based Sampling (QBS) has been proposed as a novel solution for resource estimation[2], with proceeding techniques developed therafter[3]. However, the challenge to determine if one QBS technique is better at generating resource description than another is still an unresolved issue. The initial metrics *tested* and deployed for measuring resource description quality were the Collection Term Frequency ratio (CTF) and Spearman Rank Correlation Coefficient (SRCC)[2]. The former provides an indication of the percentage of terms seen, whilst the later measures the term ranking order, although neither consider the term frequency, which is important for resource selection. We re-examine this problem and consider measuring the quality of a resource description in context to resource selection, where an estimate of the probability of a term given the resource is typically required. We believe a natural measure for comparing the estimated resource against the actual resource is the Kullback-Leibler Divergence (KL) measure. KL addresses the concerns put forward previously, by not over-representing low frequency terms, and also considering term order[2]. In this paper,

we re-assess the two previous measures alongside KL. Our preliminary investigation revealed that the former metrics display contradictory results. Whilst, KL suggested a different QBS technique than that prescribed in [2], would provide better estimates. This is a significant result, because it now remains unclear as to which technique will consistently provide better resource descriptions. The remainder of this paper details the three measures, the experimental analysis of our preliminary study and outlines our points of concern along with further research directions.

2. THE MEASURES

CTF: is a measure of the proportion of terms contained in the estimated resource description (RD_e). CTF considers the intersection of terms t in the actual resource description (RD_a), such that $CTF = \frac{\sum_{t \in RD_e} n(t, RD_a)}{\sum_{t \in RD_a} n(t, RD_a)}$, where $n(t, RD_a)$ is the number of times t occurs in RD_a. CTF captures the coverage of terms as opposed to their frequency, and was proposed to minimise bias from low frequency terms.

SRCC: accounts for the relative position of term rank shared between the actual and estimate resource vocabulary (or the intersection), by measuring the correlation (Spearman Rank) between the term rankings of RD_e and RD_a[2].

KL: measures the divergence between the probability of a t occurring in the RD_a (i.e. $p(t|RD_a)$), and the probability of a t occurring in the RD_e (i.e. $p(t|RD_e)$). Defined by $KL(RD_a||RD_e) = \sum_{t \in V} p(t|RD_a) log \frac{p(t|RD_a)}{p(t|RD_e)}$. The smaller the KL divergence score the more accurate the resource description is, with a zero KL score indicating two identical distributions. KL provides an intuitive and unambiguous measure of the resource description where, (1) the relative term frequency is captured through the probability distribution, and, (2) low frequency terms are not overly weighted because the contribution of a term to the divergence is proportional to $p(t|RD_a)$. While KL has been applied previously to this problem, it was only computed across the common intersection of terms which exist between the RD_e and RD_a[3]. We believe that evaluating the vocabulary intersection only is not appropriate for comparing competing estimates. The KL scores are directly incomparable because of the mismatch in vocabulary between sparse estimates. To account for the sparsity of the RD_e, Laplace smoothing is used to ensure a fair comparison.

3. EVALUATION

The three measures were analysed using a similar experimental approach adopted in [2]. Resource descriptions were
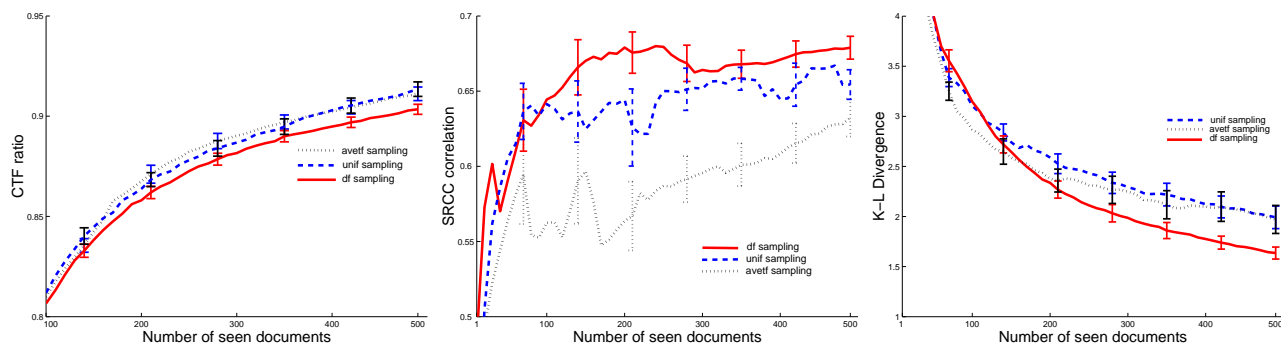
Figure 1: The change in CTF, SRCC and KL (left to right) as the number of documents sampled increases. Error bars indicate the variability across runs (shown only at various intervals). For clarity, the plot of the CTF measure only displays the results after a number of documents have been added.

estimated for each data collection using one of three QBS selection strategies, where the query terms were selected according to document frequency (df) the average term frequency (avetf), or uniformly (unif). To initialise sampling, a single query term was selected at random from an existing resource. Four documents were retrieved with each query submitted, with QBS document sampling being curtailed once 500 unique documents were seen. After each query the CTF, SRCC and KL values were recorded. The entire process was repeated 10 times per QBS method to obtain an estimate of the variance in performance measures. The experiments were performed on a number of TREC test collections, although for brevity, results are reported only for the WT2G collection (Figure 1).

The CTF ratio for each sampling method increased rapidly as more documents were sampled, eventually converging around 90% (see Figure 1). Both the unif and avetf methods obtained similar resource descriptions in terms of CTF after 500 documents were sampled. In fact, both unif and avetf generated resource descriptions that recorded significantly higher CTF ratios in comparison to df. This result would suggest the unif and avetf approaches estimated better resource description representations, however, when examining the same resource descriptions using the SRCC measure, a different trend was found. The df method obtained resource descriptions with (significantly) higher SRCC, followed by unif then avetf. This result was a reverse of the CTF findings, indicating that df obtained resource descriptions that were more highly correlated to the actual resource when compared against the other term selection strategies. An interesting observation when evaluating resource descriptions using the SRCC measure, was that as the number of documents sampled increased, many resource description estimates displayed increased variance and fluctuation (in terms of SRCC). In some cases, the mean resource description SRCC score deteriorated dramatically before increasing again. The avetf method in particular displayed many local minima, with a very sharp decrease in correlation after approximately 80 documents were sampled. In contrast, when evaluating the same resource descriptions with the KL measure, df generated resource descriptions that were significantly closer to the actual resource, with little difference between unif and avetf techniques after 500 documents. Initially, unif obtained better estimates before converging quickly, while the df method steadily improved up until 500

documents were seen. Overall, when using KL as a measure it would appear the resource descriptions improved steadily in quality as more documents were sampled.

4. DISCUSSION AND FUTURE WORK

We have argued and shown that the current measures CTF and SRCC are problematic in nature for measuring resource description quality. The application of KL provided an intuitive indication of description quality which implicity captured what CTF and SRCC were trying to measure. When using KL for comparing different QBS techniques the previously unsupported hypothesis that more frequent terms will obtain better resource description estimates was supported. This is a significant finding because much subsequent research has employed the previously accepted sampling technique[2, 4]. Further analysis is still required to provide conclusive evidence that KL is indeed a reliable indicator of resource description quality in the context of overall DIR performance. The real litmus test being whether a resource description with lower KL will result in improved resource selection accuracy. Other future research will be directed towards analysing these measures over a larger collection of differing resources, and across a number of operational settings. The final goal of this research is to achieve a better understanding of the impact of resource description quality on both resource selection and data-fusion, so that more intelligent sampling techniques may be developed.

5. REFERENCES

[1] J. P. Callan. *Advances in information retrieval*, chapter Distributed information retrieval, pages 127–150. Kluwer Academic Publishers, 2000.

[2] J. P. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, 2001.

[3] P. G. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage. In *SIGMOD'04*, pages 767–778. ACM Press, 2004.

[4] L. Si and J. P. Callan. Modeling search engine effectiveness for federated search. In *Twenty Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005.