Towards Better Measures: Evaluation of Estimated Resource Description Quality For Distributed IR

Mark Baillie Leif Azzopardi Fabio Crestani
Computer and Information Science Department
University of Strathclyde
Glasgow, United Kingdom
{mb, leif, fabioc}@cis.strath.ac.uk

Abstract

An open problem for Distributed Information Retrieval systems (DIR) is how to represent large document repositories, also known as resources, both accurately and efficiently. Obtaining resource description estimates is an important phase in DIR, especially in non-cooperative environments. Measuring the quality of an estimated resource description is a contentious issue as current measures do not provide an adequate indication of quality. In this paper, we provide an overview of these currently applied measures of resource description quality, before proposing the Kullback-Leibler (KL) divergence as an alternative. Through experimentation we illustrate the shortcomings of these past measures, whilst providing evidence that KL is a more appropriate measure of quality. When applying KL to compare different QBS algorithms, our experiments provide strong evidence in favour of a previously unsupported hypothesis originally posited in the initial Query-Based Sampling work.

1 Introduction

The acquisition and representation of an information resource still remains an unresolved issue particulary when applied to an uncooperative distributed retrieval environment (e.g. the hidden-web [5] or digital libraries [11]). When cooperation with an information resource provider cannot be guaranteed, it is necessary to obtain an unbias and accurate description of the underlying content, with respect to a number of constraints including costs (computation and monetary), consideration of intellectual property, handling legacy and non-cooperative systems and different indexing choices of the resource provider [2]. The accepted solution for resource acquisition is Query-Based Sampling (QBS) [3] and subsequent related methods [5, 7, 9]. During QBS, a sample of documents is retrieved from the underlying resource by submitting random queries to that resource. The queries are randomly selected to ensure that an unbiased resource estimate is achieved, with the varying QBS strategies differing on how the query term(s) are chosen. Sampling is terminated when it is believed that a sufficient representation of the actual resource has been obtained.

Typically, the estimated contents of a resource is represented by the distribution of terms, document frequencies, and number of documents within the resource (the estimated database size). The resource descriptions can then be utilised by both resource selection and data-fusion algorithms [2, 4, 10, 13, 15, 16]. It is therefore crucial to measure the resource description quality for a number of reasons: (1) how to determine whether a sufficiently good and unbiased resource description is obtained, (2) how to compare competing QBS algorithms, (3) how to compare competing resource descriptions, and (4) how to develop sensible termination methods based on resource description quality. The former three reasons ensure resource selection accuracy, while the later ensures efficient resource estimation. When using the currently applied measures [3, 5] to compare resource estimation techniques (and hence estimated resource descriptions), antidotal evidence exists to suggest that the findings from current measures are ambiguous and contradictory [1, 3].

The initial metrics *tested* and employed for measuring resource description quality are the Collection Term Frequency ratio (CTF) and Spearman Rank Correlation Coefficient (SRCC) [3]. The former provides an indication of the percentage of terms seen, whilst the later is an indication of term ranking order, although neither consider the term frequency, which is an important information source for all resource selection algorithms. As a consequence, both measures are required to be used in conjunction when measuring either the quality of a resource estimate, or the technique that produced the resource estimate.

While it is unclear which QBS selection method actually obtains a better resource representation, the accepted selection method is based on uniform query term sampling. However, in Callan *et al.* [3], it was originally hypothesised that selecting frequently occurring terms would be more effective at obtaining a random and unbias sample of the documents in the resource (and a better resource estimate as a consequence). It was assumed that those terms that were like stopwords would retrieve a more "random" sample of documents, minimising the likelihood of the QBS algorithm becoming trapped sampling a particular subset of the resource. This hypothesis was not supported. The current measures instead concluded that the uniform selection method was preferable and this method has been used extensively in subsequent research [3, 7, 11, 12, 13, 14].

In defence of the original hypothesis, we believe that it was not the assumption at fault but rather the measures used for evaluation. This anomaly is the underlying motivation of this work and so we reinvestigate this hypothesis. To address this issue either the two measures are required to be combined in some manner or an alternative measure employed. In this paper, we argue and show that the Kullback-Leibler divergence (KL) provides a more appropriate and natural measure for ascertaining the quality of a resource, with respect to the resource selection process, and that this hypothesis holds. Hence, the remainder of this paper is a follows. Section 2 introduces the current measure as well as KL, providing a qualitative discussion concerning the strengths and weaknesses of each measure. An empirical study of the measures is then undertaken, where in Section 3 the experimental methodology applied in the paper is outlined, and in Section 4 the results are reported. Finally, in Section 6, we conclude by summarising our findings in the context of Distributed IR, outlining future research directions.

2 Measuring Resource Description Quality

The different measures that have been applied for measuring the resource description quality are Collection Term Frequency(CTF), Spearman Rank Correlation Coefficient (SRCC) [3], and the Kullback-Leibler(KL) divergence[5]. The merits of each measure are argued in the context of measuring resource description quality.

2.1 Collection Term Frequency

To measure the quality of a resource description, it was assumed that the number of terms contained in the estimate could be used as a yard stick[3]. To avoid bias by low frequency terms, the proportion of terms was instead measured using the proposed Collection Term Frequency (CTF) ratio. The CTF ratio measures the proportion of term occurrences in the actual resource which are covered by terms in the estimated resource description. The term coverage for a given estimated resource description (RD_e) with respect to the actual resource description (RD_a) is derived by considering the intersection of terms t and their frequencies in RD_a, such that,

$$CTF = \frac{\sum_{t \in RD_e} n(t, RD_a)}{\sum_{t \in RD_a} n(t, RD_a)} \quad (1)$$

where $n(t, RD_a)$ is the number of times t occurs in RD_a.

It is assumed that as the CTF ratio approaches one, the quality of the resource description improves e.g. the coverage of terms in RD_e tends toward the actual resource RD_a. For example, if CTF ratio for a resource description is 0.8, that would indicate that the resource description accounts for 80% of all the term occurrences in the actual collection. However, CTF would be inaccurate in indicating resource quality when a document that contains a large proportion of the vocabulary is sampled. For instance, a dictionary or glossary of terms used in the resource. A resource description containing a high coverage of terms but with little knowledge of term frequency information will potentially have a negative impact on resource selection. To address this shortcoming, the Spearman Rank Correlation Coefficient was also proposed to provide additional information on the quality of the resource description.

2.2 Spearman Rank Correlation Coefficient

SRCC accounts for the relative position of the term ranking shared between the actual and estimated resource vocabulary (or the intersection), by measuring the (Spearman Rank) correlation between the term rankings within RD_e and RD_a[3]. SRCC is defined by ρ such that,

$$\rho = \frac{1 - \frac{6}{n^3 - n}(\sum d_i^2 + \frac{1}{12}\sum(f_k^3 - f_k) + \frac{1}{12}\sum(g_m^3 - g_m))}{\sqrt{(1 - \frac{\sum(f_k^3 - f_k)}{n^3 - n})}\sqrt{(1 - \frac{\sum(g_m^3 - g_m)}{n^3 - n})}} \quad (2)$$

where d_i is the ranked difference between the terms in the intersection, where the terms are ranked with respect to their document frequency values from both the actual end estimated resource. n is the total number of unique terms, f_k is the number of ties in the k^{th} group, and g_m is the number of ties in the m^{th} group. The number of ties are considered because within two vocabularies there are many terms that share the same document frequency value e.g. very infrequent terms at the tail of the vocabulary distribution.

A SRCC score closer to one would indicate high correlation between the estimated and actual resource, while a SRCC score closer to zero would suggest that no relationship exists between resource description and actual resource e.g. a poor estimate.

An initial drawback identified with measuring SRCC for resource description quality is that it becomes computationally expensive as the vocabulary in the estimate resource description increases. The other and more serious drawback is that only the rank of the terms are compared, while the frequency information is ignored. Given the potential scenario of sampling a glossary/dictionary document containing most of the terms in the resource, the SRCC score will be low because we have encountered many terms but we have not enough data to accurately estimate the correct term rank. However, the corresponding CTF score will be high due to the number of new terms now contained in the estimated resource description. Overall, SRCC would appear to be a better measure than CTF because if the terms are accurately ranked, then the frequency information may be inferred by using Zipf's law.

2.3 Kullback-Leibler

The Kullback-Leibler divergence (KL) is specifically designed for measuring the difference between two probability distributions[6]. When applied to the problem of resource description quality, KL measures the relative entropy between the probability of a term t occurring in the actual resource RD_a (i.e. $p(t|RD_a)$), and the probability of the term t occurring in the resource description RD_e (i.e. $p(t|RD_e)$). KL is defined as,

$$KL(RD_a||RD_e) = \sum_{t \in V} p(t|RD_a) log \frac{p(t|RD_a)}{p(t|RD_e)} \quad (3)$$

where,

$$p(t|RD_a) = \frac{n(t, RD_a)}{\sum_{t \in RD_a} n(t, RD_a)} \quad (4)$$

and,

$$p(t|RD_e) = \frac{\sum_{d \in RD_e} n(t, d) + \alpha}{\sum_t (\sum_{d \in RD_e} n(t, d) + \alpha)} \quad (5)$$

$n(t, d)$ is the number of times t occurs in a document d and α is a small non-zero constant (Laplace smoothing). The smaller the KL divergence the more accurate the resource description is, with a zero KL score indicating two identical distributions. To account for the sparsity within the RD_e, Laplace smoothing was applied to alleviate the zero probability problem[16] and to ensure a fair comparison between each estimated resource description (RD_e)[1]

While KL has been applied previously to the problem of resource description quality, it was only computed across the common intersection of terms that exist between the RD_e and RD_a[5]. Hence, the KL scores are not directly comparable between different resource description estimates because of the mismatch in vocabularies.

With respect to the goal of acquiring an accurate description of the resource for the selection process, the KL divergence is intuitively appealing. The probability distributions of the actual and estimated resources capture the relative (or normalised) term frequencies, which is pertinent to many of the state of the art resource selection algorithms[4, 10, 14, 16]. Apart from avoiding the need to employ two measures, KL also fulfills the criteria set forth in [3], of (1) measuring the correspondence between the estimated and actual resource vocabulary while not overly weighting low frequency terms (CTF), and also (2) measuring the correspondence between the estimated and actual frequency information (SRCC). Essentially, the KL divergence measures this phenomena precisely, while CTF and SRCC are surrogate indicators of resource description quality.

3 Experimental Methodology

The three measures were analysed using a similar experimental approach performed in [3]. Resource descriptions were 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. In[3], the initial query term was selected from a superset of one of the resources. To avoid potential bias, we used a subset of the Reuters corpora when selecting the initial seed.

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, with different randomly drawn query terms, to obtain an estimate of the variance in performance measures.

These experiments were performed on a number of TREC test collections including WSJ88-89, AP88-89, TREC-1,2,3 WT2G and WT10G. For brevity, results are reported only for the WT2G collection, which is a resource containing web-pages. However, similar trends were reported across all test collections. Figure 1 displays a summary of the results for the WT2G collection for the CTF measure. Figure 2 is a plot of SRCC, and Figure 3 is displays quality of resource description estimate measured by KL. Each plot displays the the mean measurement as further documents were sampled by the resource description estimate. At various intervals the standard error is also plotted to highlight the variability of each sampling approach.

[1] It is important to note that to compare two estimates it is required that $KL(RD_a||RD_e)$ is computed and not $KL(RD_e||RD_a)$, as the KL divergence is not symmetric.

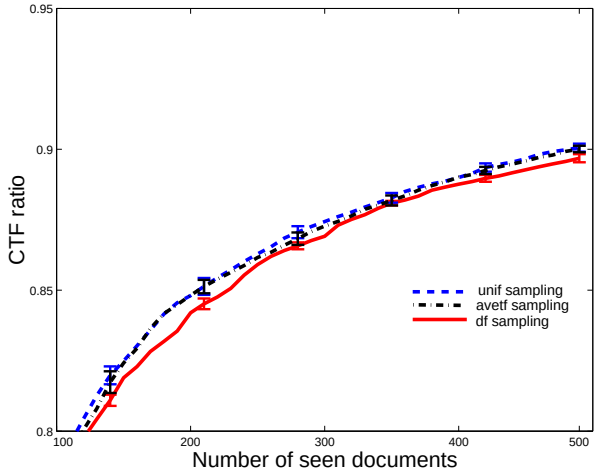

Figure 1. The change in CTF as the number of documents sampled increases. Error bars indicate the variability across runs (shown only at various intervals). For clarity, the plot of the CTF measure only displays the results after a number of documents have been added.

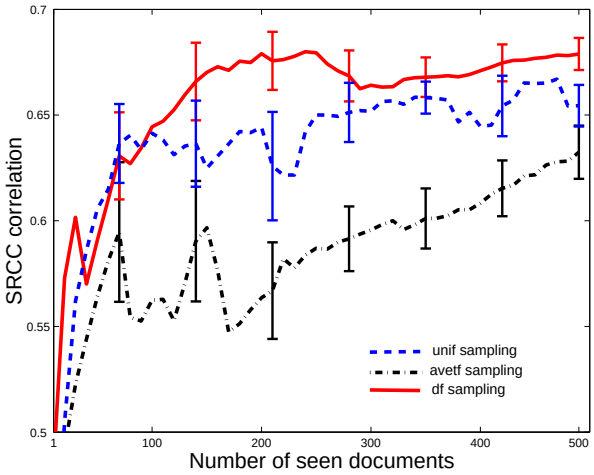

Figure 2. The change in SRCC as the number of documents sampled increases.

4 Results

The CTF ratio for each sampling method increased rapidly as more documents were sampled, eventually converging around 90% (see Figure 1). A sharp rise in CTF was found during the first 100 documents sampled across all QBS methods not shown in Figure 5. 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 (see Figure 2). 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

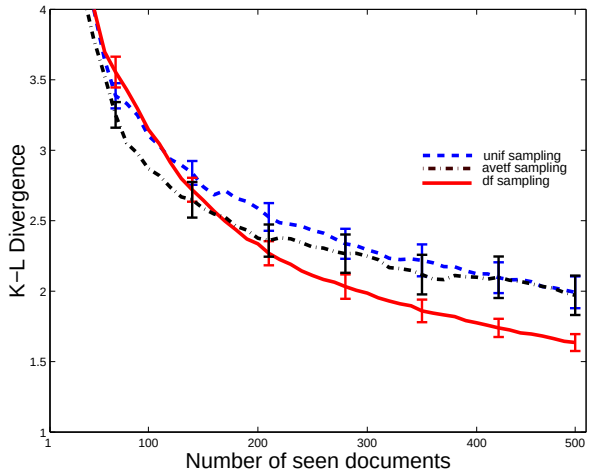

Figure 3. The change in KL as the number of documents sampled increases. For clarity, the plot of the KL measure only displays the KL score between a focused range of values.

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 (Figure 3), *df* generated resource descriptions that were significantly closer to the actual resource, with little difference between *unif* and *avetf* techniques after 500 documents. A sharp drop in KL divergence was found for all QBS techniques during the first 100 documents sampled, not shown in Figure 3. 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 (as expected).

4.1 Relationship between the measures

From the results, it would appear that if CTF is high then SRCC will be low on average, and vice versa. To examine this result in more detail, we calculated the correlation between CTF and SRCC at each query sampling, across all QBS approaches. This correlation was used for quantifying the strength of the agreement between the two measures. That is, CTF was high and SRCC was also high, then a positive correlation would exist. Alternatively, if one is high and the other is low, then we would see an negative correlation between the measures. Intuitively, we would prefer both measures to be in agreement i.e measuring the same thing. Figures 4, 5 and 6 display the results of this experiment where the x-axis represents the number of unique documents sampled, and the y-axis is the correlation between the two measures. The three lines display each term selection method: *df*, *unif* and *avetf*.

Figure 4 displays the comparison between CTF and SRCC. We note the change in the correlation as more documents were sampled. Here we can see that a negative correlation existed, confirming our hypothesis that both measures do provide contrary evidence. We posit that this result is due to the limited number of documents sampled and if we were to sample more documents, we would expect this relationship to eventually turn positive. Further work is required to determine when this state exists and whether significantly more sampling is required to obtain agreement between the measures. Otherwise, since CTF and SRCC measure different aspects of resource description quality - term coverage and ranking, then a trade off is required and it becomes unclear which QBS method is preferable. A solution would be the combination of CTF and SRCC into a single point estimate of resource description quality. Alternatively a different measure that combines both aspects could be adopted.

We also compared the relationship between KL and the two existing measures, see Figures 5 and 6. After 500 documents were sampled, there was a strong negative correlation

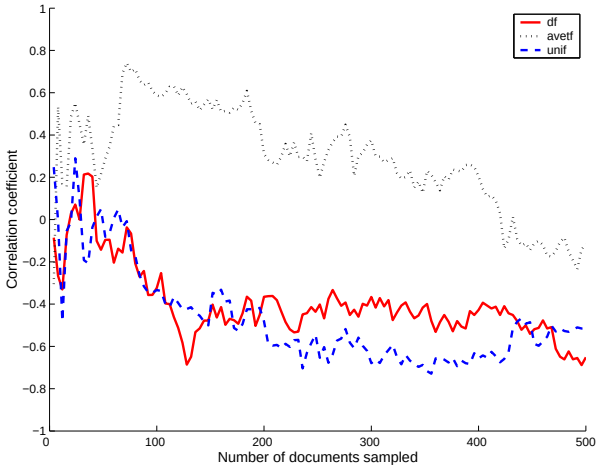

Figure 4. A plot of the relationship between the CTF and SRCC measures as more documents are sampled.

between CTF and KL (Figure 5), particulary when using *df* and *unif* as term selection strategies. This would indicate the a resource description estimate with a high CTF ratio would have a low KL score, as expected.

Conversely, there was a strong positive relationship between KL and SRCC (Figure 6). This highlighted that a resource description with a high correlation would have a poorer KL score. In other words, a resource description estimate with a similar term ranking to the actual collection, when comparing the vocabulary intersection, will invariably have a poorer KL score. The poor KL score would indicate that the term probability distribution of the estimate would be further apart from the actual collection. Intuitively, this result is not what we would expect as KL also measures the term rank (and also term frequency) between the actual and estimated resource descriptions.

5 Discussion

These initial experiments have highlighted a problem with the current metrics applied for measuring resource description quality. The results revealed that both CTF and SRCC produce contrary results. For example, both the *unif* and *avetf* techniques generated resource descriptions with significantly higher CTF ratio compared to *df*. In contrast, *df* obtained resource descriptions that had significantly stronger (SRCC) correlation to the actual (WT2G) resource. Further analysis of the relationship between CTF and SRCC identified that the measures were in fact negatively correlated (after 500 documents were sampled). This indicated that more often than not, a higher CTF ratio for a resource description estimate would result in lower SRCC.

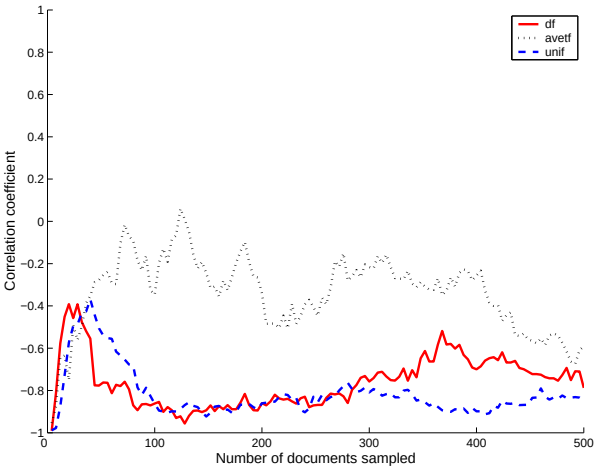

Figure 5. A plot of the relationship between the CTF and KL measures as more documents are sampled.

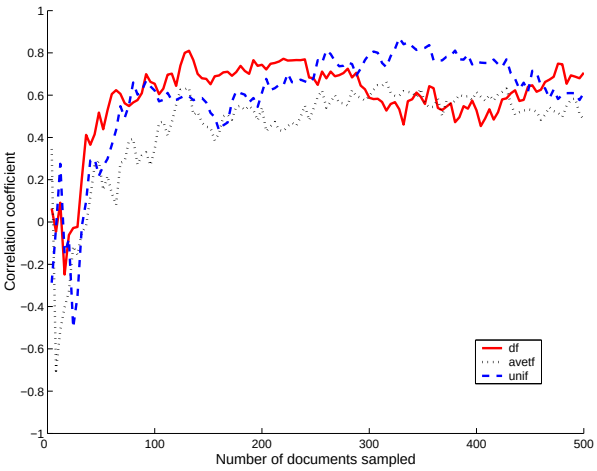

Figure 6. A plot of the relationship between the SRCC and KL measures as more documents are sampled.

If both measures are an indication of resource description quality, then this result would imply that during QBS, a trade off has to be considered between high CTF and strong SRCC correlation. In other words, a middle ground would have to be found between balancing high CTF ratio and SRCC correlation to find the 'optimal' resource description. Such a result raises doubts about the validity of employing both metrics together when measuring resource description quality, providing evidence to support our initial concerns.

When analysing both in isolation, there were also doubts on the validity of either measure. For example, CTF measures the coverage of terms in the resource description compared to the actual resource. It is not clear if high coverage of terms does indeed reflect resource description quality, especially as term frequency information is not reflected by CTF. A resource description covering a high number of terms, but without no indication as to term contribution may be meaningless, as many terms will not contribute to the content stored in a resource[8]. A resource description with a poor estimate of the actual term frequencies could be chosen over a better resource description that has worse term coverage. Such a scenario would have a negative impact on the overall quality of the DIR system.

Focusing on the SRCC measure, it was discovered that all QBS methods produced highly variable estimates. As further documents were seen, the correlation between the estimate and the actual resource fluctuated instead of rising steadily as expected. A possible reason for this fluctuation could be the vocabulary intersection. Only the correlation between the common terms shared by the actual and estimated resource description is measured with SRCC. Therefore, as more documents are seen by a resource description, new unseen terms are also included, increasing the vocabulary intersection as a consequence. When a large number of content rich documents containing new unique terms are added, the term ranking accuracy initially deteriorates. The correlation only improves again when further documents are added which provide a enough information to correctly re-rank the terms rankings, which in turn is reflected by an increase in SRCC. This evidence raises doubts on calculating the SRCC of the vocabulary intersection, with a followup analysis required to determine if this is a true measure of description quality, or most likely, a problematic metrics.

Alternatively, the KL divergence provides a more appropriate and natural measure for ascertaining the quality of a resource by measuring the difference between the estimated and actual resource term distribution. Since, the term distribution is exactly what resource selection algorithms use in the selection process. It also fulfills the criteria set forth in [3] for an adequate measure, that is: (1) measuring the correspondence between the estimated and actual resource vocabulary while not overly weighting low frequency terms,

and also (2) measuring the correspondence between the estimated and actual frequency information. A key finding when analysing resource descriptions with KL, was that the original QBS hypothesis of selecting more frequent occurring (*df*) query terms from the estimated resource description would obtain better, unbiased representations of the resource.

As expected a strong relationship was found between CTF and KL, where an estimate with a high CTF score would have a low KL score. This result indicated, that those resource description estimate with high term coverage, would also have a closer term probability distribution to the actual collection. This we assume is an indication of the quality of the resource description estimate obtained through QBS. Conversely, the relationship between SRCC and KL was the reverse of what we expected indicating potential problems with analysing only the vocabulary intersection between the actual and estimated resource description.

Finally, a common limitation of all three measures employed in this work, is that they all require that the actual collection description to be known, *a priori*. This is not problematic in the development phase of DIR, where full control is possible. However, in an operational setting such measures are impractical, because this information is unavailable. This highlights the need to identify techniques for assessing resource quality that can be employed without recourse to the actual resource description. This is especially so in environments where the resources are dynamic and as a consequence resource descriptions are continually required to be updated. For example, a measure of the quality of the resource would help notify the DIR system when resources are required to be updated, and also when to stop QBS sampling. For a number of reasons, intelligent mechanisms for QBS termination are desirable over the currently applied stopping criteria that are based on heuristics. QBS is expensive in terms of time and resources, so only the optimal number of documents for accurate resource selection and data fusion should be sampled. Also, by applying threshold based rules such as a maximum number of unique documents seen, cannot generalise to varying collections (in terms of the number of documents, length and type of documents, etc). Therefore, adaptive approaches based on resource description quality, which are not reliant on prior resource knowledge, are required to maximise QBS while minimising wastage. The development of such a technique would provide a solution to this unresolved problem and is left to future work.

6 Conclusions 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[3, 7, 11, 12, 13, 14]. 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 there is a correlation between KL and resource selection accuracy. Future research will be directed in this direction in order 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.

7 Acknowledgements

This research was undertaken as part of the PENG project. PENG, “PErsonalised News content programminG Information”, is a Specific Targeted Research Project (IST-004597) funded within the Sixth Program Framework of the European Research Area. The authors would also like to thank Murat Yakici for his helpful comments throughout this work.

References

[1] M. Baillie, L. Azzopardi, and F. Crestani. An evaluation of resource description quality measures. In *The 21st Annual ACM Symposium on Applied Computing (ACM SAC 2006)*, Dijon, France, April 23-27 2006. ACM.

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

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

[4] L. Gravano, H. García-Molina, and A. Tomasic. GlOSS: text-source discovery over the Internet. *ACM Transactions on Database Systems*, 24(2):229–264, 1999.

[5] 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.

[6] S. Kullback. Information theoery and statistics. *Wiley, New York*, 1959.

[7] J. Lu and J. P. Callan. Pruning long documents for distributed information retreival. In *CIKM’02*, pages 332–339. ACM Press, November 4-9 2002.

[8] H. P. Luhn. Automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[9] G. Monroe, J. French, and A. Powell. Obtaining language models of web collections using query-based sampling techniques. In *HICSS'02-Volume 3*, Washington, DC, USA, 2002. IEEE Computer Society.

[10] H. Nottelmann and N. Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 290–297, New York, NY, USA, 2003. ACM Press.

[11] H. Nottelmann and N. Fuhr. The mind architecture for heterogeneous multimedia federated digital libraries. In *ACM SIGIR 2003 Workshop on Distributed Information Retrieval*, Canada, 2003. ACM.

[12] P. Ogilvie and J. P. Callan. Experiments using the lemur toolkit. In *Text REtrieval Conference*, 2001.

[13] L. Si and J. P. Callan. Unified utility maximization framework for resource selection. In *CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management*, pages 32–41, New York, NY, USA, 2004. ACM Press.

[14] 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.

[15] L. Si, R. Jin, J. P. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 391–397, New York, NY, USA, 2002. ACM Press.

[16] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 254–261, New York, NY, USA, 1999. ACM Press.