Adaptive Query-Based Sampling of Distributed Collections

Mark Baillie, Leif Azzopardi, and Fabio Crestani

Department of Computing and Information Sciences,
University of Strathclyde, Glasgow, UK
{mb, leif, fabioc}@cis.strath.ac.uk

Abstract. As part of a Distributed Information Retrieval system a description of each remote information resource, archive or repository is usually stored centrally in order to facilitate resource selection. The acquisition of precise resource descriptions is therefore an important phase in Distributed Information Retrieval, as the quality of such representations will impact on selection accuracy, and ultimately retrieval performance. While Query-Based Sampling is currently used for content discovery of uncooperative resources, the application of this technique is dependent upon heuristic guidelines to determine when a sufficiently accurate representation of each remote resource has been obtained. In this paper we address this shortcoming by using the Predictive Likelihood to provide both an indication of the quality of an acquired resource description estimate, and when a sufficiently good representation of a resource has been obtained during Query-Based Sampling.

1 Introduction

An open problem that Distributed Information Retrieval systems (DIR) face is how to represent large document repositories, also known as resources, both accurately and efficiently. To facilitate resource selection, the process of assessing which collections contain relevant information with respect to a user's information request, a description of each information resource a DIR service searches is required. The obtained resource descriptions form a collection selection index that enables the DIR system to determine which online collections to search given a query [6]. Therefore, obtaining precise resource descriptions is an important phase as the quality of such representations will impact on resource selection accuracy, and ultimately retrieval performance. The acquisition and representation of an information resource presents many research challenges, particularly in uncooperative environments. When co-operation from an information resource provider cannot be guaranteed, it is necessary to obtain an unbiased 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 different indexing choices of the resource provider [6, 11]. While Query-Based Sampling is currently used for content discovery of uncooperative resources, the application of this technique

is dependent upon heuristic guidelines to determine when a sufficiently accurate representation of each remote resource has been obtained. In this paper we address this shortcoming by using the Predictive Likelihood to provide both an indication of: (i) the quality of an acquired resource description estimate, and (ii) when a sufficiently good representation of a resource has been obtained during Query-Based Sampling.

The remainder of this paper is structured as follows. First, we provide a brief outline of Query-Based Sampling and how it can be used to build resource descriptions, then we outline how Predictive Likelihood can be adopted as a measure of resource description quality with respect to the user information needs (Section 2). Next, we compare Predictive Likelihood to existing measures and show that it provides a comparable indication of resource quality despite the fact no *a priori* knowledge is used (Section 3). Finally, we demonstrate and evaluate the application of Predictive Likelihood in Query-Based Sampling on two DIR testbeds (Section 4). Our analysis validates that this unsupervised approach can substantially reduce the number of documents sampled without detracting from resource selection accuracy. We then conclude the paper with a short discussion detailing the implications of using such an approach and indicate directions for future work (Section 5).

2 Query-Based Sampling and Predictive Likelihood

A resource description is a representation of the content contained within a resource (e.g. a document collection). It can take a variety of forms depending on a number of influencing factors; such as the retrieval model used for resource selection, and the level of co-operation between a search service and information provider. Currently adopted representations include a term vector of counts or probabilities (i.e. a language model) [7], a sample of indexed documents from each collection [14], or indeed the full index [6].

The widely accepted solution for resource description acquisition is Query-Based Sampling (QBS) [7]. During QBS an estimated representation is obtained by submitting random queries to the actual collection, incrementally adding the newly retrieved documents to the estimated resource representation. Queries are randomly selected to ensure that an unbiased resource estimate is achieved. Sampling is then terminated when it is believed a sufficiently good representation of the underlying resource has been acquired, facilitating effective retrieval. Through empirical analysis, the number of documents required to be sampled, on *average*, was estimated to be approximately 300-500. This was believed to obtained a sufficiently good representation of a resource [7]. This threshold was estimated by measuring the estimated resource description against the actual resource using two indicators of quality, and then considering the corresponding retrieval selection accuracy.

While it has been shown that this criterion provides adequate resource selection accuracy under certain conditions, there are potential limitations. A fixed threshold will not always generalise across other collections and environments.

Cases when the blanket application of such a heuristic would be inappropriate include: (i) when the sizes of resources are highly skewed, and (ii) when the resources are heterogeneous. In the former, if a resource is large then undersampling may occur because not enough documents are obtained. Conversely, if a collection is small in size, then oversampling may occur, increasing costs beyond necessity. In the latter case, if the resource is varied and highly heterogeneous then to obtain a sufficiently accurate description would require more documents to be sampled than when a resource is homogenous. For both scenarios, adopting a threshold based heuristic will not ensure a sufficiently good resource description for all resources. This has been recently verified by Shokouhi *et al.* [13] over a number of different DIR testbeds.

Ideally QBS should be curtailed only when a sufficiently good description of the resource has been acquired such that the number of documents sampled is minimised and system performance preserved. In this paper we argue that the Predictive Likelihood of the user's information needs given the estimated resource description can be utilised as a measure of the goodness of a resource description estimate. We believe that the Predictive Likelihood can be used to: (i) provide an indication of the resource description quality, and (ii) to indicate when a sufficiently good representation of the resource has been obtained.

In statistical modelling the log-likelihood of a model on a held out sample of data is often applied as a measure for the "goodness of fit" of that model. This measure is also known as the Predictive Likelihood (PL) of the model [8]. PL is generally used to measure the quality of a language model in the fields of Statistical Language Modelling, but has been more recently applied to estimate language model parameters in text retrieval [1, 10, 16]. In these studies it has been generally assumed that those models which maximise PL will achieve better retrieval performance. Following this intuition, in the context of measuring description quality, we aim to maximise the PL of the user's information needs given the estimated resource description. By using PL we are measuring how representative each distributed information resource is when compared to the known (typical) information needs of the users of the DIR system. This is a departure from the original QBS assumption that a resource description should be a sufficient sample of the actual entire collection. Instead, by using PL descriptions are measured with respect to the information needs of the users of the system. Before discussing this main difference, we first define the Predictive Likelihood measure and how it incorporates the user's information needs.

Formally, given a sequence of queries $Q = \{q_{ij} : 1, \ldots, N; 1, \ldots, M\}$, where q_{ij} is the j^{th} term of the i^{th} query, which corresponds to a particular term t in the estimated resource description $p(t = q_{ij}|\hat{\theta})$. The likelihood of a resource description estimate $\hat{\theta}$ generating Q is given by the conditional probability:

$$p(Q|\hat{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(t = q_{ij}|\hat{\theta})$$

where,

$$p(t|\hat{\theta}) = \frac{n(t,\hat{\theta})}{\sum_{t' \in \hat{\theta}} n(t',\hat{\theta})}$$

and $n(t,\hat{\theta})$ is the number of times term t occurs in the resource estimate $\hat{\theta}$. We engage the standard assumption of independence between query terms and also between queries [16]. For computational connivence, however, we use the Predictive Log Likelihood of the estimated resource $\hat{\theta}$:

$$\ell(\hat{\theta}, Q) = \log p(Q|\hat{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(t = q_{ij}|\hat{\theta})$$

Using this approach for measuring the quality of a resource description is fundamentally different to existing standard approaches. Current methods measure the quality of an estimate against the actual resource, thus requiring full collection knowledge *a priori*. As mentioned previously such information is not readily available except in artificial or simulated environments. In comparison, PL requires that a set of queries Q are available for evaluating each resource description instead of the actual collection. Therefore, the selection of this set of queries is an important step in training the DIR system.

We assume that the set of queries Q are representative of the information needs of the users of that system. To elaborate, these information needs, or queries, can take the form of: (i) the previous interactions of the system obtained through the query logs [2] or (ii) profiles that represent the interests of the user-base, similar to profiles used in Information Filtering systems [4]. In the former, a query set consistent with the information needs of the user-base of the system can be obtained from query logs. For instance, the query logs of each user can be mined to extract a representative set of queries. Alternatively, if no historical queries are freely available, it is possible to access example queries from Information Retrieval test collections or a similar web based corpus. Conversely, or even supplementary, users of the system could be profiled explicitly, such as through a questionnaire or survey, where profiles represent typical topics, subject areas and tasks that the users of the system will undertake. However, both solutions for representing Q enable the DIR system to be tuned either towards an *average* user-base or even tailored towards specific users or user groups depending on the requirements of the system. Throughout the development of the system, Q can also be re-assessed with respect to the user's dynamically changing needs.

3 Predictive Likelihood as an Indicator of Quality

In this section PL is evaluated and compared as a measure of resource description quality alongside a currently adopted method. In this experiment we are motivated to evaluate whether a relationship exists between PL and the currently applied measure. If a relationship does exist, this will provide evidence that PL can be utilised as a surrogate measure of resource description quality

with the added advantage that PL does not require *a priori* knowledge of the underlying information resource statistics.

3.1 Existing Measures of Resource Description quality

Current measures of resource description quality include the Collection Term Frequency ratio (CTF), Spearman Rank Correlation Coefficient (SRCC) [7], and the Kullback-Leibler (KL) divergence [3, 11]. CTF and SRCC are normally applied in tandem, where the former provides an indication of the percentage of terms seen, while the latter is an indication of term ranking order, although neither consider the term frequency which is an important information source for all resource selection algorithms. In a recent study, the SRCC measure was shown to be unstable and unreliable [3]. As an alternative measure the KL divergence was proposed. With respect to the goal of measuring the quality of a resource description the KL divergence is appealing for a number of reasons. The term probability distributions of the actual and estimated resource descriptions capture the relative (or normalised) term frequencies, when an accurate estimation of such information is pertinent to many of the state of the art resource selection algorithms [6, 11, 14, 15]. It also fulfils the criteria set forth in the original QBS study by Callan and Connell [7] of measuring the correspondence between the estimated and actual resource vocabulary while not overly weighting low frequency terms (CTF), and also measuring the correspondence between the estimated and actual frequency information (SRCC). Essentially the KL divergence measures this phenomena precisely, resulting in a more stable and precise measure in comparison to the surrogate indicators CTF and SRCC.

We therefore compare the KL against the PL. In this experiment we hypothesise that the PL will provide a comparable indication of the resource description quality to KL.

3.2 Kullback-Leibler divergence

The Kullback-Leibler Divergence (KL) provides a measure for comparing the difference between two probability distributions[12]. 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 θ (i.e. $p(t|\theta)$), and the probability of the term t occurring in the resource description $\hat{\theta}$, i.e. $p(t|\hat{\theta})$. Formally, the KL Divergence is defined as:

$$KL(\theta|\hat{\theta}) = \sum_{t \in V} p(t|\theta) log \frac{p(t|\theta)}{p(t|\hat{\theta})}$$

where, $p(t|\theta) = \frac{n(t,\theta)}{\sum_{t \in \theta} n(t,\theta)}$, $p(t|\hat{\theta}) = \frac{\sum_{d \in \hat{\theta}} n(t,d)+\alpha}{\sum_{t}(\sum_{d \in \hat{\theta}} n(t,d)+\alpha)}$, $n(t,d)$ 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 description, with a score of zero indicating two identical distributions. To account for the

Collection	# Documents	# Collection Terms	# Unique Terms	Mean Doc. Length
Aquaint	1,033,461	284,597,335	707,778	275
WT10g	1,692,096	675,181,452	4,716,811	399

Table 1. Collection Statistics.

sparsity within the set of sampled documents, Laplace smoothing is applied to alleviate the zero probability problem and to ensure a fair comparison between each estimated resource description.

3.3 Experimental Methodology

Our aim is to evaluate whether PL provides a similar indication of the *true quality* of a resource description estimate. Here, we assume that the KL divergence is the true measure of quality because it's measurement is taken against the actual resource description (ground truth). Our hypothesis is that for a set of estimated resource descriptions the PL measure will rank these estimated resource descriptions in the same order as the KL measure. If this is the case then PL will provide a comparable indication of the quality of that resource according to the KL measure.

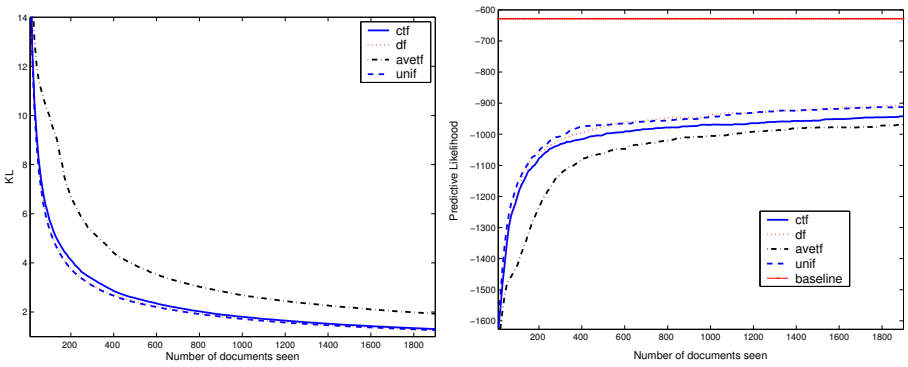

Fig. 1. Measuring the quality of resource description estimates obtained from Aquaint collection by the four QBS approaches. KL and PL measurements for each sampling approach are displayed as the number of documents sampled increases.

The experiments were performed on several different TREC collections, with varying characteristics. For brevity, though, we only report on two of these collections, the news collection Aquaint, and the Web collection WT10g (See Table 1).

Estimated resource descriptions were then created for these collections using QBS as follows:

1. A term is randomly selected from an unrelated vocabulary and is used as the first query for sampling.

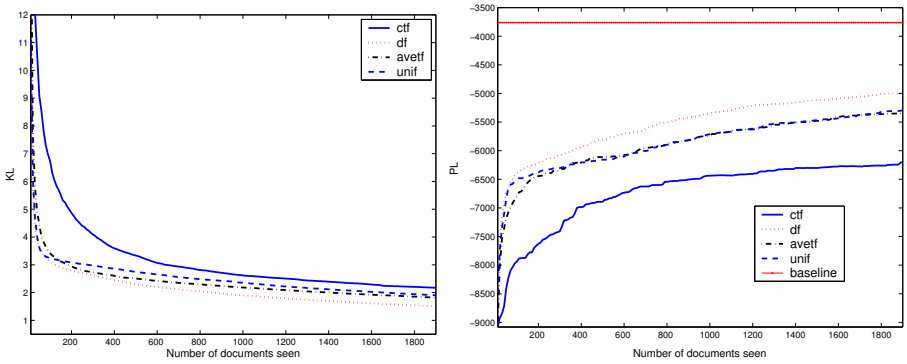

Fig. 2. Measuring resource description estimates obtained from the WT10g collection.

2. The resource is queried and the top four documents returned are added to the estimated resource description.
3. The KL and PL are measured and recorded.
4. The next query is generated using the currently estimated resource description using one of the four sampling strategies: the collection frequency (*ctf*), the document frequency (*df*), the average term frequency (*avetf*), or randomly (*unif*) [7].
5. If the stopping criterion has not been satisfied, return to step (2).

We continued sampling until we obtained 2000 documents. For each sampling strategy the entire process was repeated 25 times because the initial term affects the quality of the resource description. This generated 100 estimated resource descriptions for each collection along with the corresponding measurements. The query set to compute PL for both collections consisted of TREC Topics 1-200. The title field from these topics were extracted as queries which formed Q for each collection respectively.

3.4 Experimental Results

Resource description quality Figures 1 and 2 summarise the performance of each sampling strategy by displaying the mean quality score over the 25 runs for the Aquaint and WT10g collection respectively. In the KL plots, a score of zero indicates that the estimate is identical to the actual description. While in the PL plots, the higher the PL the better the quality, where the baseline is shown as a solid line which denotes the PL score for Q given the actual resource description θ.

We were first concerned with the rate of improvement as more documents were added to each resource description estimate. The general trend when measuring the quality of resource descriptions, as further documents were sampled, appeared to be similar (See Figures 1 and 2). As the number of documents initially sampled increased, a sharp drop in KL and a corresponding rise in PL,

	Documents Sampled			
Collection	**200**	**500**	**1000**	**2000**
Aquaint	0.85*	0.68*	0.62*	0.53*
WT10g	0.38*	0.51*	0.57*	0.82*

Table 2. The Kendall τ Correlation of KL and PL for ranking resource description estimates in terms of quality, recorded at different intervals of documents sampled. An asterisk indicates a statistically significant correlation at $p < 0.05$.

was found. As QBS sampling continued, the rate of improvement for each resource description levelled out using either measure. This trend indicated that by adding further documents to the estimate provided small gains in quality. At this point, a decision to terminate QBS based on the cost of sampling further documents versus the gain in further representation of the resource should be made. For the Aquaint collection, Figure 1, this point occurs when approximately 800-1200 documents are sampled across all term selection methods. For the WT10g collection, Figure 2, this was found at approximately 1200-1600 for KL and somewhat later when measuring with PL.

We were also concerned with which term selection method (i.e. *df, unif, ctf* or *avetf*) acquired the better resource description estimates in terms of KL and PL, and in particular if there was agreement between both measures. Focusing first on the Aquaint collection, the ordering of the mean quality of each sampling strategy was found to be identical when using KL and PL. For both measures the rank order was: *unif*, *df*, *ctf* then *avetf* (best to worst). For the WT10g collection, both measures also ranked the methods the same: *df*, *unif*, *avetf* followed by *ctf*.

Both KL and PL ranked the resource descriptions obtained from each term selection method in the same order. However, across the two collections this rank order varied with the random term selection method (*unif*) preferred for Aquaint, while the document frequency strategy (*df*) considered better for WT10g. This is an unexpected outcome as it reveals that PL can be used in a novel way for determining which sampling method will provide a better estimate on a per collection basis, potentially increasing sampling effectiveness during QBS.

Correlation between the measures We ranked all the estimated resource descriptions, irrespective of term selection strategy, according to KL and PL producing two ranked lists. We then compared the ranked lists produced by each measure using Kendall's τ correlation test at various points in the sampling process. By doing so, we could determine if there was a strong concordance between the rankings (i.e. quantify how close in agreement each measure is when ranking the different resource description estimates in terms of quality). This approach has been used previously in IR to compare different measures of retrieval performance in [5]. The assumption is that a good estimate would be ranked highly for both measures. A correlation score close to 1 would indicate that two measures have identical rankings. A score closer to 0 would indicate no relationship

between the measures. Table 2 provides the τ correlation coefficient at 200, 500, 1000 and 2000 documents sampled.

At each of the different sampling points, shown in Table 2, the results reveal that there was a close agreement between both ranked lists complied using both measures. This relationship was found to be statistically significant across both collections, and at each of the different intervals, providing stronger evidence to support our hypothesis that the PL measure provides a comparable indication of quality with respect to the KL measure.

4 Predictive Likelihood as a Stopping Criterion

QBS is an iterative process where sampling is curtailed when a single or set of stopping criteria has been reached. In the standard approach to QBS, once n unique documents have been retrieved then sampling is stopped [7]. We propose to use the PL measure to inform the decision making process in order to decide when enough documents have been sampled. Our stopping criterion is based on the difference in the PL score for the estimated resource description, between the previous iteration $k-1$ and the current iteration k of the sampling process. The difference ϕ_k at iteration k, where $k > 1$, is given by:

$$\phi_k = \ell(\hat{\theta}_k, Q) - \ell(\hat{\theta}_{k-1}, Q) = log\left(\frac{p(Q|\hat{\theta}_k)}{p(Q|\hat{\theta}_{k-1})}\right)$$

where $\hat{\theta_k}$ is the resource description estimate at the k^{th} iteration. If ϕ_k is below a threshold ϵ, then sampling is curtailed, where ϵ indicates the necessary amount of improvement required to continue sampling. By doing so we are using a gradient ascent optimisation to maximise the Predictive Likelihood of the estimated resource description given Q [9]. The ratio of PL scores provides an indication of the rate of improvement over the previous iteration. Consequently, the free parameter ϵ is independent of the document collection characteristics (such as size and heterogeneity). Unlike the fixed n document curtailment strategy, this parameter is generalisable to other collections.

By using this technique we believe that a sufficiently good estimation of the resource will be obtained, which will minimise any unnecessary wastage from oversampling, and will also avoid obtaining an insufficient sample through undersampling. We further hypothesise that because sufficient representations of each resource will be obtained, this will translate into better selection accuracy over the fixed method. We shall now refer to the proposed method as QBS-PL and the previous threshold based approach as QBS-T.

4.1 Evaluation

The aim of the next set of experiments was to determine whether QBS-PL provided better resource selection accuracy over QBS-T. This was examined in two ways: (1) if QBS-PL improved selection accuracy when the number of sampled

documents were approximately equal, and (2) if QBS-PL provided comparable resource selection accuracy to QBS-T when the number of sampled documents were substantially less than the threshold approach.

Experimental Settings Two DIR testbeds based on the TREC Aquaint collection were formed for these experiments, with the documents partitioned By-source and By-topic. The By-source testbed contains 112 simulated collections, with the documents arranged into collections based on both the news agency that published each document, and the month the document was published. In this testbed the size of each collection is uniform. The By-topic testbed contains 88 collections, with documents grouped by topical similarity using single pass *k-means* clustering. In this testbed, collection sizes are skewed and represent a realistic setting with respect to the distribution of content.

For QBS, sampling was performed with the term selection strategy set to *df*, with four documents retrieved per query. The thresholds used for the QBS-T ranged from 100-1000 unique documents. For QBS-PL, ϵ was set to 0.01. We also include descriptions using the full collection information ('complete') as a benchmark (i.e. all the documents in the resource to build the description). To provide an indication of how sensitive the retrieval accuracy is when applying QBS-PL with different query sets, we used four different sized query sets Q constructed from 200 TREC Topics (Topics 1-200). The number of queries in each set were 50, 100, 150 and 200. So as not to train and test using the same set of queries, another set of queries from the TREC HARD 2005 track were used for resource selection. This set contained 50 test topics, where the title field was used as the query. Resource selection was performed using the DIR benchmark algorithm CORI [6]. Resource selection accuracy was measured using the recall-based $\hat{R}$ metric. $\hat{R}$ is a measure of the overall percentage of relevant documents contained in the top r collections [7]. We measured $\hat{R}$ at $r = \{5\%, 10\%, 15\%, 20\%, 25\%\}$ of all collections searched. We also captured the average number of documents sampled per collection, and the total number of documents overall.

Experimental Results Table 3 provides an overview of the results obtained for QBS-PL, QBS-T and also using the complete estimates (not all thresholds for QBS-T are shown). Figure 3 is a plot of QBS-PL using 200 queries, QBS with two document thresholds ($t = 500, 1000$), and the resource selection performance using complete information, compared across each of the $\hat{R}@r$ values.

The performance of the QBS-PL method varied as the size of the Q increased, see Table 3. For both collections, an increase in the Q coincided with an improvement in $\hat{R}$, across all collection cut-offs, with the QBS-PL $Q = 200$ method performing best over both testbeds. In both cases, there was a steady increase in the number of documents sampled as the size of Q grew. We suspect that this would tail off as more queries are added, but this is as yet unconfirmed due to the finite number of test queries available. As more queries are added to Q, it is sensible to expect more documents will be sampled in order to cover the new subject areas expressed in these queries. This is intuitively appealing

Aquaint: By-source testbed							
Parameters	$\hat{R}$@5%	$\hat{R}$@10%	$\hat{R}$@15%	$\hat{R}$@20%	$\hat{R}$@25%	Ave. docs.	Total docs.
QBS-PL, $Q = 50$	0.093	0.162	0.231	0.283	0.341	247.9	27767
QBS-PL, $Q = 100$	0.110	0.182	0.248	0.301	0.366	347.3	38893
QBS-PL, $Q = 150$	0.116	0.188	0.250	0.308	0.360	434.8	48699
QBS-PL, $Q = 200$	**0.126**	**0.212**	**0.279**	**0.332**	**0.378**	**500.6**	**56066**
QBS-T $n = 300$	0.108	0.179	0.248	0.308	0.360	300	36960
QBS-T $n = 500$	**0.113**	**0.191**	**0.249**	**0.310**	**0.362**	**500**	**56000**
QBS-T $n = 1000$	0.124	0.207	0.291	0.353	0.415	1000	112000
Complete	0.163	0.249	0.315	0.390	0.454	11743.9	1033461
Aquaint: By-topic testbed							
QBS-PL, $Q = 50$	0.63	0.73	0.79	0.83	0.85	235.2	20466
QBS-PL, $Q = 100$	0.64	0.74	0.82	0.84	0.87	344.2	29952
QBS-PL, $Q = 150$	0.65	0.74	0.82	0.85	0.87	394.4	34315
QBS-PL, $Q = 200$	**0.66**	**0.75**	**0.82**	**0.85**	**0.86**	**456.4**	**39685**
QBS-T $n = 500$	0.54	0.69	0.75	0.81	0.84	500	44000
QBS-T $n = 1000$	0.58	0.73	0.79	0.84	0.87	1000	88000
Complete	**0.64**	**0.74**	**0.81**	**0.85**	**0.88**	**2262**	**1033461**

Table 3. Each technique is evaluated by $\hat{R}@r$ percent of the collections searched, and the overall document statistics for each QBS technique across the two testbeds.

because as the information needs of the users of a system diversify and change, a larger number of documents will be required in order to sufficiently describe a resource given those needs.

In comparison with the threshold method, QBS-T, the performance of QBS-PL provides comparable selection accuracy while reducing the number of documents sampled (See Figure 3 and Table 3). If we consider QBS-PL $Q = 200$ on the By-source testbed, the fixed threshold of $n = 500$ returns a similar number of documents sampled, but QBS-T's selection accuracy is worse. It is not until the threshold was increased to $n = 1000$ that similar selection accuracy was obtained. However, this means that over 55,000 extra documents were sampled, an increase of almost 100%. On the By-topic testbed, QBS-PL provides better accuracy when compared with the QBS-T estimates. Even when QBS-T was set to $n = 1000$, with an increase of 40,000 to 50,000 extra documents sampled over the QBS-PL estimates, the selection accuracy was still 6-12% worse. It was only when complete information was used that performance similar to QBS-PL $Q = 200$ was obtained. The seems to suggest that there are problems with under and over sampling of many collections, which was not so problematic when collection size was uniform as in the By-source testbed (and in previous work [7]); but is problematic when the collection sizes are skewed.

5 Conclusions and Future Work

Both experiments have shown that PL can be effectively used as a measure of resource description quality, and as a consequence can be integrated into the

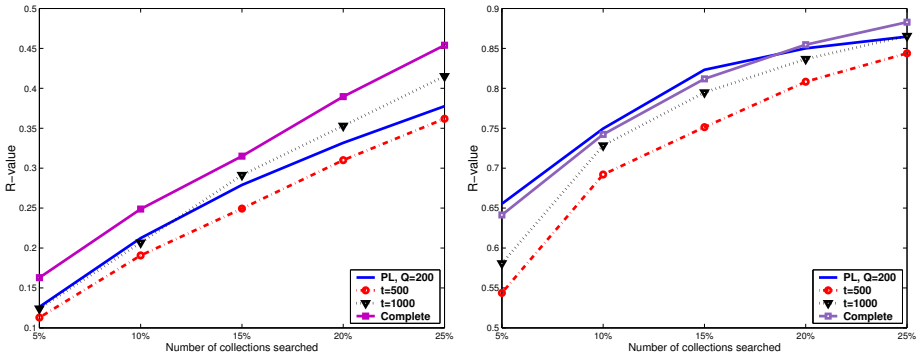

Fig. 3. QBS-PL versus QBS-T across a range of cut off values of r over By-source (left) and By-topic (right) respectively.

QBS algorithm. It was shown that a significant relationship exists between PL and KL divergence. However, PL is a radical departure from existing measures such as KL. It is radical because it questions whether a completely unbiased representation of the underlying resource is actually required. By using PL, we are not measuring quality in terms of sampling a sufficiently good representation of the actual collection, but measuring whether the resource description estimate satisfies the information needs of the users of the DIR system. With PL we measure each resource description with respect to a set of queries that represent the typical information needs, Q, of the user-base of a system i.e. evaluating each estimate with respect to the information users want from that resource. For example, by increasing Q, it was highlighted that further documents were required to be sampled before a sufficient representation of the collection was obtained. This increase in the number of documents required to satisfy Q mirrored the addition of new information needs in Q.

We then highlighted that the problem of under and oversampling does exist when employing a QBS algorithm which uses a fixed document threshold (QBS-T). As previously posited, this is especially problematic in a situation when resources are of varying size and content. Consequently, the efficiency and effectiveness of the QBS approach is compromised when using such criteria. By employing QBS-PL, it was shown that this problem can be addressed. Using PL to measure resource description quality without *a priori* knowledge of each distributed collection, the original QBS algorithm was improved both in terms of accuracy and efficiency. QBS-PL minimised the problems of under and oversampling, and in particular when faced with collections of varying size and content, we were able to determine when a sufficiently good representation of each collection had been obtained, which in turn was reflected by performance gains. In contrast, a fixed threshold resulted both in poorer resource selection performance and also increased overheads.

A main advantage of utilising PL is that it enables the resource descriptions to be tailored specifically to the information needs of the user. This is appealing

and paves the way for the development of personalised (distributed) retrieval systems. Defining the query set Q provides an intuitive mechanism for obtaining resource descriptions that are personalised to specific users or user groups; an unexplored area of research that we are currently investigating.

6 Acknowledgements

This work was supported by PENG, a Specific Targeted Research Project funded within the 6th PF of the European Research Area. More information on the project can be found at http://www.peng-project.org/.

References

1. L. Azzopardi, M. Girolami, and C. J. Risjbergen. Investigating the relationship between language model perplexity and IR precision-recall measures. In *Proceedings of the 26th ACM SIGIR conference*, pages 369–370, 2003.
2. R. A. Baeza-Yates. Applications of web query mining. In *Proceedings of the 27th ECIR*, pages 7–22, Santiago de Compostela, Spain, 2005.
3. M. Baillie, L. Azzopardi, and F. Crestani. Towards better measures: Evaluation of estimated resource description quality for distributed IR. In *First International Conference on Scalable Information Systems*. IEEE CS Society, 2006.
4. N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin. *Communications of the ACM*, 35(12):29–38, 1992.
5. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd ACM SIGIR conference*, pages 33–40, 2000.
6. J. P. Callan. *Advances in information retrieval*, chapter Distributed information retrieval, pages 127–150. Kluwer Academic Publishers, 2000.
7. J. P. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions of Information Systems*, 19(2):97–130, 2001.
8. M. H. Degroot. *Optimal Statistical Decisions (Wiley Classics Library)*. Wiley-Interscience, April 2004.
9. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.
10. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.
11. P. G. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage. In *Proceedings of the ACM SIGMOD Conference*, pages 767–778, 2004.
12. S. Kullback. Information theoery and statistics. *Wiley, New York*, 1959.
13. M. Shokouhi, F. Scholer, and J. Zobel. Sample sizes for query probing in uncooperative distributed information retrieval. In *APWeb 2006*, volume 3841, pages 63–75. Springer Lecture Notes in Computer Science, 2006.
14. L. Si and J. P. Callan. Modeling search engine effectiveness for federated search. In *Proceedings of the 28th ACM SIGIR Conference*, pages 83–90, 2005.
15. J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd ACM SIGIR conference*, pages 254–261, 1999.
16. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transaction of Information Systems*, 22(2):179–214, 2004.