Adaptive Query-Based Sampling For Distributed IR

Leif Azzopardi
CIS Department
University of Strathclyde
Glasgow, UK
leif@cis.strath.ac.uk

Mark Baillie
CIS Department
University of Strathclyde
Glasgow, UK
mb@cis.strath.ac.uk

Fabio Crestani
CIS Department
University of Strathclyde
Glasgow, UK
fabioc@cis.strath.ac.uk

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Selection Process

General Terms: Algorithms, Experimentation

Keywords: Distributed Information Retrieval, Query-Based Sampling, Selection Accuracy, Efficiency

1. INTRODUCTION

In Distributed Information Retrieval systems (DIR), the widely accepted solution for resource description acquisition is Query-Based Sampling (QBS) [1]. In the standard approach to QBS, once 300-500 unique documents have been retrieved sampling is curtailed. This threshold was obtained by empirically measuring the estimated resource description against the actual resource, and then considering the corresponding retrieval selection accuracy [1]. However, a fixed threshold may not generalise to other collections and environments beyond that which it was estimated on (i.e. a set of resources of uniform size [1]). Cases when the blanket application of such a heuristic would be inappropriate include (1) when the sizes of resource are highly skewed and (2) when the resources are very heterogenous. In the former, if a resource is very large then undersampling will occur because not enough documents were obtained. Conversely, if a collection is very small in size, then oversampling will occur increasing costs beyond necessity. In the later 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 resources are homogenous. Either way, adopting a flat cut off will not necessarily provide sufficiently good resource descriptions for all resources.

In this poster we propose an adaptive QBS technique which samples until a sufficiently good description of the resource has been obtained according to the past information needs of users. In an evaluation against the threshold based approach, we show that both sampling efficiency and resource selection can be improved.

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.

2. ADAPTIVE QUERY BASED SAMPLING

In statistical modelling the log-likelihood of a model on a held out sample of data is often used as a measure for the "goodness of fit" of that model, also known as the Predictive Likelihood (PL) [2]. We propose to use PL to measure how representative each distributed information resource is when compared to the typical information needs of the users of the DIR system; defined by a record of their past queries. When the PL of the resource has been maximised, QBS is curtailed. 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, corresponding to a particular term t in the estimated resource description $p(t = q_{ij}|\hat{\theta})$. The predictive (log-)likelihood of the estimated resource description $\hat{\theta_k}$ generating Q at the k^{th} iteration of QBS sampling is given by the log conditional probability, expressed as (note: we assume independence between terms and queries):

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

where $p(t = q_{ij}|\hat{\theta_k})$ is the probability of term t occurring in the $\hat{\theta_k}$. To decide whether sampling should be curtailed, the difference ϕ_k at iteration k, where $k > 1$ is given by $\phi_k = \ell(\hat{\theta_k}, Q) - \ell(\hat{\theta_{k-1}}, Q)$, is used. If ϕ_k is below a threshold ϵ, then sampling is curtailed, where ϵ indicates the necessary amount of improvement required to continue sampling. This is an example of a gradient ascent optimisation [3] , where we maximise PL given Q. Importantly the free parameter ϵ is independent of the document collection characteristics such as size and heterogeneity. By using this technique, we believe that a sufficiently good estimation of the resource will be obtained, which minimises any unnecessary wastage from oversampling, and also avoids obtaining a insufficient sample through under-sampling.

One requirement of using PL is a set of queries which adequately represent the future needs (we assume that past queries will reflect this). While access to past queries is not problematic, as they can be sourced from query logs, it is an open question as to how much influence they will have on the curtailment process. Further, using queries introduces a personalised aspect to the creation and estimation of resource descriptions. This distinction leads to the possibility of personalised resource descriptions. Where, over time, as user needs change, the resource description estimates can be re-assessed to determine whether further sampling is required to satisfied this change. This is a move away from assuming a static and fixed representation of a resource to

a more dynamic, flexible and customisable representation.

3. EVALUATION

The aim of this experiment is to compare the performance of the estimates obtained by adaptive method (QBS-PL) versus threshold method (QBS-T) across resource selection accuracy, sampling efficiency and final retrieval performance. This was performed on two DIR testbeds based on the Aquaint news collection where the collection was partitioned By-source and By-topic. The By-source testbed contained 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, and similar to that used in [1]. The By-topic test bed contained 88 collections, with documents grouped by topical similarity using a single pass *k-means* clustering algorithm. When indexed, the collections were stemmed and common stop words removed. For QBS, sampling was preformed using the document frequency query term selection strategy [1]. The first four documents retrieved for a query were added to the estimated resource description [1]. The thresholds used for the QBS-T ranged from 100-1000 unique documents. For QBS-PL, ϵ was set to 0.01, and the queries used to measure the PL were the titles of the TREC Topics 1-200 (previously used on other new collections). As a baseline we also included descriptions based on the full collection information ("complete"). The DIR benchmark algorithm CORI was used for resource selection and data fusion [1] using the TREC HARD-Robust 2005 Topics (containing 50 queries).

4. RESULTS AND DISCUSSION

Table 1 provides an overview of the results obtained for QBS-PL, QBS-T and using the complete estimates (not all thresholds for QBS-T are shown for brevity). Resource selection accuracy was measured using the recall-based $\hat{R}$@r% metric, which is a measure of the overall percentage of relevant documents contained in the top r% collections [1]. Resource sampling efficiency was captured by the average and total number of documents sampled per collection. Final retrieval accuracy was measured in *R-precision* and *b-pref* (not shown in table).

In comparison with the QBS-T method, we can see that the performance of QBS-PL provides comparable selection resource accuracy while reducing the number of documents sampled. If we consider QBS-T on the By-source testbed, we find that a fixed threshold of $n = 500$ returns a similar number of documents sampled, but QBS-PL's selection accuracy is better. It is not until the threshold is increased to $n = 1000$ that similar selection accuracy is obtained by QBS-T. This requires over 55,000 extra documents to be sampled, an increase of almost 100%. On the By-topic testbed, QBS-PL provides excellent selection accuracy when compared with the QBS-T estimates. While the threshold of $n = 300$ provided comparable selection accuracy in the By-source testbed, here the result of under sampling is clearly seen with a complete degradation in performance. Even when the threshold was increased to $n = 1000$, selection accuracy was still poorer than QBS-PL, despite again requiring approximately 50,000 extra documents. It is only when full information is used for all collections do we achieve similar performance to QBS-PL.

Despite selection accuracy improvements, we found that the retrieval accuracy in terms of both *b-pref* or *R-precision* was similar (except in the case of the By-topic QBS-T $n = 300$), with no notable differences between those tested. These results can be explained by the fact that in the standard data fusion process, the top r% of collections are chosen with an equal number of documents retrieved from each collection. As a consequence any improvements attained in selection accuracy are not capitalised during the fusion and final retrieval stage (as done in [4]). We posit that employing such a technique would translate selection accuracy gains into improved retrieval effectiveness.

Aquaint: By-source testbed			
Parameters	$\hat{R}$@10%	$\hat{R}$@20%	Avg. (Total) docs.
QBS-PL	0.212	0.332	501 (56066)
QBS-T $n = 300$	0.179	0.308	300 (36960)
QBS-T $n = 500$	0.191	0.310	500 (56000)
QBS-T $n = 1000$	0.207	0.353	1000 (112000)
Complete	0.249	0.390	11744 (1033461)
Aquaint: By-topic testbed			
Parameters	$\hat{R}$@10%	$\hat{R}$@20%	Avg. (Total) docs.
QBS-PL	0.755	0.856	456 (39685)
QBS-T $n = 300$	0.227	0.495	300 (26400)
QBS-T $n = 500$	0.692	0.808	500 (44000)
QBS-T $n = 1000$	0.733	0.842	1000 (88000)
Complete	0.746	0.854	2262 (1033461)

Table 1: Results of each QBS technique.

5. CONCLUSIONS

Our preliminary findings show that the adaptive strategy QBS-PL has benefits in terms of both sampling efficiency and selection accuracy when compared to the threshold based stopping method. QBS-PL minimised overheads while maintaining resource selection performance. When faced with a situation where collection sizes were skewed, QBS-PL was found not only to be more efficient but also more effective than applying a document threshold, substantially reducing the total number of documents to be sampled, while improving recall during resource selection. Although this did not translate into better retrieval accuracy using the benchmark fusion technique (CORI) for the reasons stated above, we are encouraged to perform further research. Not only will this be directed towards using better fusion strategies, but also towards investigating the influence of using different queries when building the resource description estimates, and how these can be tailored to create personalised DIR systems.

6. REFERENCES

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

[2] M. H. Degroot. *Optimal Statistical Decisions (Wiley Classics Library)*. Wiley-Interscience, April 2004.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[4] N. Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM Trans. Inf. Syst.*, 17(3):229–249, 1999.