Sample Sizes for Query Probing in Uncooperative Distributed Information Retrieval

Milad Shokouhi, Falk Scholer, and Justin Zobel

School of Computer Science and Information Technology,
RMIT University, Melbourne 3001, Australia
{milad, fscholer, jz}@cs.rmit.edu.au

Abstract. The goal of distributed information retrieval is to support effective searching over multiple document collections. For efficiency, queries should be routed to only those collections that are likely to contain relevant documents, so it is necessary to first obtain information about the content of the target collections. In an uncooperative environment, query probing — where randomly-chosen queries are used to retrieve a sample of the documents and thus of the lexicon — has been proposed as a technique for estimating statistical term distributions. In this paper we rebut the claim that a sample of 300 documents is sufficient to provide good coverage of collection terms. We propose a novel sampling strategy and experimentally demonstrate that sample size needs to vary from collection to collection, that our methods achieve good coverage based on variable-sized samples, and that we can use the results of a probe to determine when to stop sampling.

1 Introduction

Information retrieval (IR) systems such as search engines receive queries from users, and aim to provide the most relevant information available in their databases in response. Search engines can use a central index for retrieval, but this strategy has several drawbacks. Due to hardware limitations it may not be easy to keep all the documents indexed on a single machine. Also, a centralized search engine for web data relies on documents being provided by a crawl, and thus cannot index the hidden web. For example, the query "wireless and network" returns 28013 answers (as of May 27, 2004) from the USPTO (the US Patent and Trendmark database, `patft.upsto.gov/netahtml/search-adv.htm`), while Google (`google.com`) reports no answer for searching for the same keywords in the same site [Ipeirotis, 2004].

Distributed information retrieval (DIR) addresses such problems. In DIR systems, information is held in separate collections, which might be in different physical locations or on separate *servers*. The query is first passed to a central *broker*. The broker then sends this query to all or some of the servers. The servers provide the broker with their best answers, which the broker merges into a single list that is returned to the user. Thus the documents do not have to be gathered into a single location, and the constraints imposed by machine capacity are much

X. Zhou et al. (Eds.): APWeb 2006, LNCS 3841, pp. 63–75, 2006.

more relaxed; however, DIR does introduce query-time costs of networking, and a query may be sent to collections where there are unlikely to be answers.

DIR systems therefore need to address two major issues, how to select collections and how to merge the results returned from each collection. In this paper, we investigate the first of these: which collections should be selected for each query? Brokers typically compare each query to summaries — also called representation sets — of each collection [Ipeirotis and Gravano, 2004], and choose the collections whose summaries have the greatest similarity to the query. In most previous work, and in this paper, each summary contains statistics about the lexicon of the corresponding collection. If the lexicon of the collections is provided to the central broker —that is, if the servers are *cooperative* — then complete and accurate information can be used for collection selection. In an uncooperative environment such as the hidden web, however, the collections need to be probed to establish a sample of their topic coverage. This technique is known as query probing. In previous work, it has been claimed that a probe that returns 300 documents is sufficient to characterize a collection. However, we dispute this claim.

We propose an adaptive query probing technique that uses statistics of term occurrences in returned documents to examine whether further probing is justified. Our results show that, with only 300 documents, coverage of the lexicon is small and query effectiveness is impaired. By use of larger samples, and by use of our thresholding technique that determines when sampling can terminate, we obtain much greater effectiveness. While the number of documents that must be probed is substantially increased, the method is free of an arbitrary choice of cut-off and is expected to adapt to collections with different characteristics.

2 Related Work

In a cooperative DIR environment, servers provide the broker with global information about their collections. This information is usually about the terms they contain [Callan et al., 1995; Gravano et al., 1999; Yuwono and Lee, 1997] or the similarity function they use for the ranking. The advantage of this type of environment is that broker usually has comprehensive knowledge about each collection, allowing relatively accurate selection. However, many servers do not provide such information. Approaches to cooperative DIR differ in terms of the type of information that is provided and in the merging functions that are used. In spite of implementation differences, the performance these methods are reported to be similar [D'Souza et al., 2004a;b].

In an environment such as the Web, collections are usually non-cooperative and do not publish their index information. In non-cooperative environments, brokers try to obtain information about the collections that are to be searched by sending them artificial queries and evaluating the returned answers. These queries are known as *probes*, and the whole procedure is usually called *query-probing* [Craswell et al., 2000] or *query-based sampling* [Callan and Connell, 2001]. For example, suppose that the series of single-term probe queries "soccer", "basketball",

"health", "computer", "IBM", and "cancer" has been sent to a collection and the following numbers of answers have been returned: 1500, 1730, 200, 0, 0, and 2. Then the collection is more likely to include coverage of sports than compared to computer science.

Callan and Connell [2001] applied query-based sampling for iteratively discovering the language model of the collections in non-cooperative environments. Their algorithm starts by selecting an initial query that returns at least one answer from the collection, and then retrieves the first N results returned. The language model is updated according to the new terms found in the retrieved documents. The next probe queries are selected from the obtained language model; probing continues until a *stopping criterion* is met. Callan et al. tested different values of N and various stopping conditions, and reported that using $N = 4$ and 75 queries, thus obtaining about 300 documents, leads to a good summary of the sampled collection. They also examined different strategies for query selection and conclude that these do not have a significant effect on final performance. Variants explored included choosing the queries from the terms that have the highest document frequency, collection frequency, and average term frequency in the current language model, with randomly generated queries as a baseline. In all these cases, the reported results are similar, with random queries having small advantage over other methods.

Using random queries is now a widely accepted method for query-based sampling [Gravano et al., 2003; Callan and Connell, 2001; Craswell et al., 2000]. We also use this strategy in our proposed approach. The main focus of our work is to examine different stopping criteria.

Callan and Connell [2001] proposed use of the *ctf ratio*, representing the fraction of term occurrences in the total collection that are covered by distinct terms in the sampled documents. For example, consider a collection that includes only two documents. The first consists of 98 occurrences of the term "car" and one occurrence of a single term, "book". The other document consists of a single term, "car". By finding the second document, the sampled language model will contain 99% of term occurences in the collection and the *ctf ratio* will be 0.99. Callan et al. [1999] report that after sampling about 300 documents the *ctf ratio* becomes smooth. Later, we examine the effectiveness of using the *ctf ratio* for estimating the quality of the obtained language model.

Craswell et al. [2000] investigated query probing for server selection on the web. They used the sampling approach of Callan and Connell [2001] to estimate the server effectiveness, and used this estimation for server selection. They report that a system that chooses the top 10 collections out of 956, based on summaries obtained by query probing, can outperform a central index that has indexed 25% of the total documents. Query-based sampling has recently been applied for different purposes, such as estimating the size of uncooperative collections [Si and Callan, 2003] or classification of *hidden web* databases [Gravano et al., 2003]. In all of these experiments, fixed sample sizes were used. Ipeirotis and Gravano [2004] used query expansion techniques to overcome the poor quality of the samples but their investigation was limited to topical collections.

Table 1. Collection Statistics

Collections	Number of Documents	Number of Unique Terms
UDC-1	17,352	82,434
UDC-2	17,352	83,992
DATELINE 325	16,248	82,707
DATELINE 509	30,507	106,644
WEB	301,681	2,021,973

3 Measuring the Effectiveness of Query Probing

As can be seen from the work surveyed above, most of the prior research uses fixed parameters in query based sampling, and there is no clear stopping condition and termination point for the process. To our knowledge, none of the proposed methods for non-cooperative DIR involves an adaptive choice of stopping point. Yet the usual stopping point seems low; 300 documents seem unlikely to be sufficient for representing many current collections, such as digital libraries. Intuitively, larger collections with diverse topics need more samples while smaller, topic-specific ones might need less. Williams and Zobel [2005] have shown that vocabulary growth after indexing about 45 GB of web data does not converge to zero, and, the rate of discovery of new unique terms is about one in every 400 term occurrences. For query-based sampling, the question is, therefore, when to stop sampling. In the following section we explore the following hypotheses:

1. As long as we keep sampling, the vocabulary continues to grow.
2. The rate of vocabulary growth is not a good way to estimate collection size.
3. The risk of missing significant terms is high with traditional sampling.

Test Environment. We tested query-based sampling on five collections of different sizes and contents, shown in Table 1. The first two collections are from the UDC-39 testbed (discussed in detail in the next section), each containing 17,352 documents of TREC newswire data. DATELINE 509 and DATELINE 325 are two managed collections used by D'Souza et al. [2004a]. Documents in each collection of this testbed are TREC newswire data split by the ⟨DATELINE⟩ field. They reported that gathering the data in managed collections improves the overall performance of document retrieval from distributed collections. Since documents in these collections are usually from the same organization and authors, we would expect them to have a more limited vocabulary compared to collections of similar size with various authors. The fifth collection is composed of 2 GB of data from a 1997 web crawl (the first two gigabytes of TREC WT10g collection). The WT10g collection was constructed to be representative of the web [Bailey et al., 2003].

To evaluate our approach we also extracted the most *significant* terms from each collection by gathering all terms with a Cosine $tf \cdot idf$ [Baeza-Yates and Ribeiro-Neto(1999)] factor greater than a certain threshold (γ) using Zettair.[1]

[1] Available from `http://www.seg.rmit.edu.au`

We used 0.5 for γ; other values of γ do not affect the approach and could be used. This information is used after termination of query-based sampling, as a measure of the effectiveness of the collected summaries and of the risk of missing significant terms.

Two testbeds are used in our selection experiments. *SYM236* includes 236 collections of varying size. It has been used in previous related work [French et al., 1999; Powell and French, 2003]. It includes collections made from documents on TREC CDs 1 to 4. *UDC39* is made from UDC236 [French et al., 1999; Powell and French, 2003], which contains 236 collections that include the same number of documents each and has been constructed from the same documents as SYM236 (from TREC CDs 1 to 4). The difference is only the methodology used to assign documents to collections. In SYM236, collections are of varying sizes, while in UDC236 collections contain the similar number of documents. UDC39 has 39 collections each made from concatenating six consecutive collections in UDC236. Therefore, UDC-1 in this testbed contains documents in the first six collections of UDC236. The total numbers of documents in both testbeds are the same.

We used the titles of TREC topics $51-150$ as queries, whose average length is $2-3$ terms, which is similar to web queries [Jansen et al., 2000]. A thousand answer documents are retrieved in response to each query from each selected collection. The assumption is that collections only return a limited number of documents for any given query. If a collection does not return a relevant document in the top 1000 results, the DIR system can never use that document. We used Lemur[2] for query-based sampling, and CORI [Callan et al., 1995] for collection selection and result merging because, although it may not be the most effective method, our results can then be directly compared to those in most previous work. We leave the testing of other collection selection methods, such as those which are discussed in Meng et al. [2002] as future work. For each collection (other than WEB) we gathered samples of different sizes, from 100 to 3000 documents. Each sample n contains all of the documents from sample $n-1$, plus 100 new documents. The initial sample always extracts 100 distinct documents. At each point, the system calculates the number of unique and significant terms available in the samples. We show results for 3000 documents because this number was sufficient according to our experiments; other collections might need greater sample sizes to meet the stopping criteria. We use a recall metric to measure completeness of the sampled term set:

$$Recall(s,\gamma) = \frac{\text{Number of significant terms in the sample}}{\text{Total number of significant terms}}$$

$$Term = \begin{cases} Significant & \text{if } tf \cdot idf \geq \gamma \\ Not\ significant & \text{otherwise} \end{cases}$$

Experimental Results. Figure 1 shows the number of unique and significant terms in each sample provided by query-based sampling UDC-1 and UDC-2 collections.

[2] `http://www.lemurproject.org`

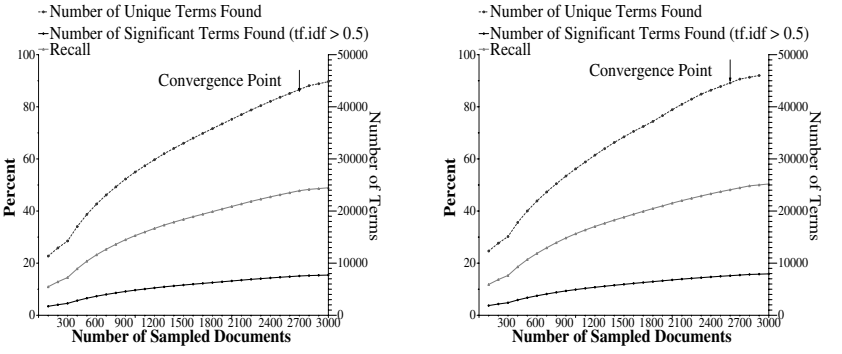

Fig. 1. Recall and total numbers of distinct terms for samples of the UDC-1 (left) and UDC-2 (right) collections

The rate at which new unique terms are found slows as the number of sampled documents increases. The slope of each curve is large at 300 documents, the recommended size for query-based sampling [Callan et al., 1999]. As sampling continues, the slope becomes flatter. Based on previous work [Williams and Zobel, 2005], continued sampling will always continue to find new words but the rate will decrease. Note that the rate for significant terms drops more rapidly than for terms.

A key contribution in this paper is that convergence to a low rate of vocabulary increase is indicative of good coverage of vocabulary by the sampled documents. In other words, query sampling reaches a good coverage of the collection vocabulary when the slope becomes less than a certain threshold; empirical tests of this hypothesis are discussed below. In these charts, when the trends for the number of unique terms starts smoothing, the curves for the number of significant terms found are nearly flat, which means that by continuing sampling we are unlikely to receive many new significant terms, and it is unlikely to be efficient to keep probing. The recall curve confirms that the number of new significant terms hardly increases after sampling a certain amount of documents. The recall value for a sample of 300 document is less than 15%, while for summaries including more than 2000 documents this amount is greater than 45% (three times more) in both graphs. These trends strongly indicate that a sample size of 300 documents is insufficient for making effective summaries. As the slopes for significant terms are not negligible after sampling 300 documents, the risk of losing significant terms is high at this point. Figure 2 shows similar trends for the DATELINE managed collections. Again, the samples made from 300 documents do not appear to be a good representation of the collection language model. Curiously, although we were expecting the graphs to get smooth sooner than the previous collections (because of the documents should have similar topics), the results are very similar. The reason might be that all the collections so far are based on the TREC newswire data and contain similar documents. Trends for discovery of new terms and recall values for summaries obtained by sampling our WEB collection are shown in Figure 3. As the collection is significantly larger, we

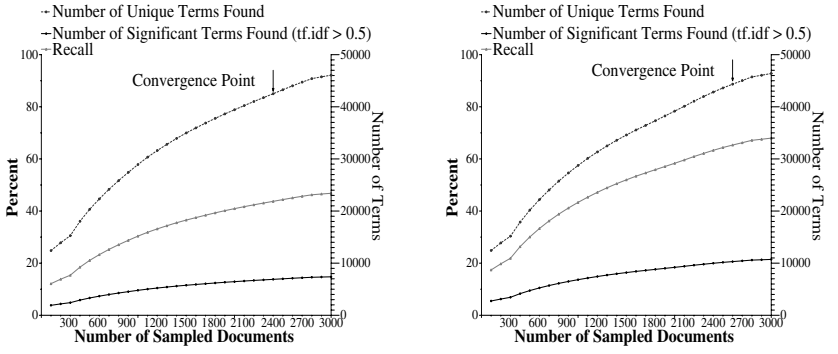

Fig. 2. Recall and total numbers of distinct terms for samples of the DATELINE 325 (left) and 509 (right) collections

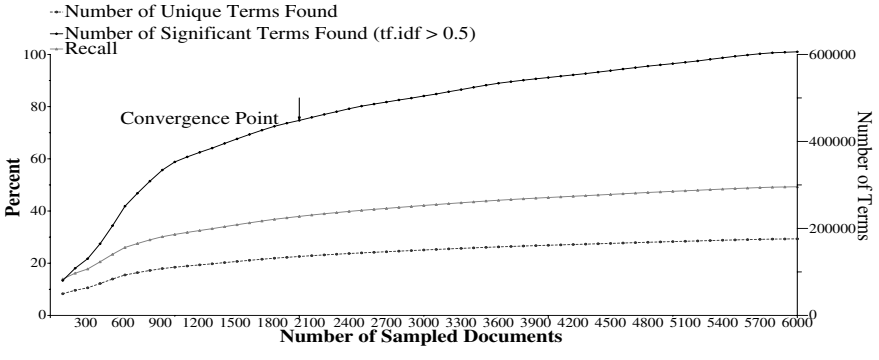

Fig. 3. Recall and total numbers of distinct terms for samples of the WEB collection

extended our range of sampling to 6000 documents. The slope is sharply upward, not only after sampling 300 documents, but also in all the other points lower than 1000. At this point, the curve for significant terms is already fairly smooth. In other words, we are unlikely to receive significant terms with the previous rate by continuing probing. Interestingly, while the system has downloaded less than 2% of total documents, the trend for discovering new terms is getting smooth. Recall values start converging after downloading nearly 900 documents. Based on these experiments, we conclude that:

- Hypothesis 1 is clearly confirmed, since the accumulation of new vocabulary never stops completely.
- Hypothesis 2 is confirmed, because collections that were significantly different size show similar rates of vocabulary growth. For example DATELINE 325 and DATELINE 509 produced similar trends, although they are very different in size.
- Hypothesis 3 is confirmed; if probing is halted after sampling 300 documents, the risk of losing significant terms is high.

4 Distributed Retrieval with Variable-Sized Samples

Given that a sample size of 300 is inadequate, but that some condition is needed to terminate sampling, we need to investigate when sampling should cease. In this section, we test the effect of varying the sample size on retrieval effectiveness. Table 2 shows the mean average precision (MAP) for different sample sizes. We use the TITLE field of TREC topics $51 - 150$ as queries. Values for precision at 10 and 20 documents retrieved are provided because these include the documents that users are most likely to look at [Jansen et al., 2000]. Cutoff values represent the number of collections that will be searched for each query. The results show that, by using samples of more than 300 documents, the overall performance increases. The previously recommended number of 300 documents is not in general a sufficient sample size. Previous work uses *ctf* as an indication of vocabulary coverage, and shows that curves become smooth after downloading a limited number of documents from a collection [Callan et al., 1999; Callan and Connell, 2001]. However, our results show *ctf* is not an indication of achieving good vocabulary coverage. Terms that are more frequent in the collection are more likely to be extracted by query probing. Once the system finds such a term, the *ctf ratio* increases more than when system finds a word with lower frequency. However, these terms are not necessarily more important than the other terms [Luhn, 1958] in the collection, and indeed are unlikely to be significant in queries; downloading them does not mean that the coverage of the vocabulary is sufficient. Given that 300 documents is insufficient, and that the appropriate number is not consistent from collection to collection, the question is: how big a sample should be chosen from a given collection?

We propose that an appropriate method is to keep sampling until the rate of occurrence of new unique terms (the slope in previous figures) becomes less than a predefined threshold. Specifically, we propose that query probing stop when, for η subsequent samples, the rate of growth in vocabulary becomes less than a threshhold τ. Based on the empirical experiments discussed in the previous

Table 2. The impact of changing sample size on effectiveness

Testbed	Summary Size	Cutoff	MAP	P@10	P@20
SYM236	300	10	0.0133	0.1465	0.1256
SYM236	700	10	0.0370	0.2765	0.2474
SYM236	900	10	0.0326	0.2510	0.2260
SYM236	300	20	0.0222	0.1616	0.1506
SYM236	700	20	0.0533	0.2806	0.2587
SYM236	900	20	0.0506	0.2888	0.2536
UDC39	300	10	0.0611	0.2653	0.2566
UDC39	900	10	0.0739	0.2878	0.2724
UDC39	1500	10	0.0773	0.2959	0.2867
UDC39	300	20	0.0881	0.2949	0.2765
UDC39	900	20	0.0972	0.3051	0.2867
UDC39	1500	20	0.1016	0.2969	0.2878

Table 3. Effectiveness of a central index of all documents of SYM236 or UDC39

Relevant Retrieved	MAP	P@10	P@20	R-Precision
8776	0.1137	0.2939	0.2760	0.1749

Table 4. Effectiveness of two DIR systems using both samples of 300 documents and adaptive sample sizes, for SYM236 ($\eta = 3$, $\tau = 2\%$)

Cutoff	Relevant Retrieved	MAP	P@10	P@20	R-Precision
Samples of 300 documents					
1	158	0.0023	0.0682	0.0435	0.0063
10	1396	0.0133	0.1465	0.1256	0.0429
20	2252	0.0222	0.1616	0.1506	0.0616
50	3713	0.0383	0.1628	0.1676	0.0926
118	4800	0.0515	0.1430	0.1395	0.1032
Adaptive samples					
1	527	0.0075	0.1454	0.1244	0.0168
10	2956	0.0327	0.2510	0.2199	0.0772
20	4715	0.0532^{**}	0.2724	0.2372	0.1135^{*}
50	6813	0.0823^{**}	0.2796^{**}	0.2633^{**}	0.1506^{**}
118	7778	0.0936^{**}	0.2388^{**}	0.2327^{**}	0.1604^{**}

section, we suggest initial parameter choices of $\eta = 3$ and $\tau = 2\%$; that is, probing stops once three consecutive probes all show growth rate of less than 2%. These *convergence* points are indicated by arrows in previous figures. In our approach, these points indicate when sampling is "enough". According to the observations, "enough" varies drastically from collection to collection. Increasing the value for η or decreasing τ delay reaching the stopping condition and increase the number of samples that should be gathered from the collection.

SYM236. The performance of a central index for document retrieval for both collections is shown in Table 3. Since both testbeds include exactly the same documents, the central index for both of them is the same. We used the values in this table as the baseline. Central indexes are usually reported as being more effective than distributed systems [Craswell et al., 2000]. The first column is the number of relevant documents retrieved for TREC topics $51 - 150$; the last column is the precision of the system after as many documents have been retrieved as there are relevant documents in the collection. A comparison of the effectiveness of two systems using traditional and adaptive query-based sampling techniques is shown in Table 4. The numbers above the middle line represent the values obtained from the traditional method, while those below specify the same factor using our adaptive method. For *cutoff* $= 1$, only the best collection — that whose sampled lexicon has the greatest similarity to the query — will be searched. For *cutoff* $= 118$, half of the collections will be searched. It can be seen that our method outperforms the traditional query probing technique

Table 5. Summary of sampling for SYM236 and UDC39, using adaptive and traditional sampling

Testbed	Method	Documents	Unique Terms	Min	Max
SYM236	Traditional (300 documents)	37,200	831,849	300	300
SYM236	Adaptive ($\tau = 2\%,\ \eta = 3$)	163,900	1,565,193	500	2700
SYM236	Adaptive ($\tau = 1\%,\ \eta = 3$)	321,300	2,083,700	500	3200
UDC39	Traditional (300 documents)	11,700	624,765	300	300
UDC39	Adaptive ($\tau = 2\%,\ \eta = 3$)	80,800	1,289,607	1400	2800

Table 6. Effectiveness of two DIR systems using both samples of 300 documents and adaptive sample sizes, for UDC39 ($\eta = 3$, $\tau = 2\%$)

Cutoff	Relevant Retrieved	MAP	P@10	P@20	R-Precision
Samples of 300 documents					
1	1132	0.0161	0.2061	0.1658	0.0351
10	5551	0.0611	0.2653	0.2566	0.1273
20	7320	0.0881	0.2949	0.2765	0.1610
30	7947	0.0969	0.2735	0.2622	0.1705
Adaptive samples					
1	1306	0.0178	0.2173	0.1699	0.0403*
10	6342	0.0764**	0.2959**	0.2837**	0.1465**
20	7826	0.1017**	0.3051	0.2969**	0.1730**
30	8280	0.1089**	0.3051**	0.2837**	0.1790**

in all of the parameters and for all cutoff values [3]. Sanderson and Zobel [2005] demonstrated that a significant improvement in performance requires statistical tests. We applied the t-test for comparing the outputs of traditional and adaptive systems. Values shown with an asterisk (*) are significantly different at $P < 0.05$ while those with double asterisks (**) differ significantly at $P < 0.01$.

Table 5 gives more information about the number of terms and documents that have been sampled using the traditional and adaptive techniques. The smallest and largest samples in each testbed are specified in the last two columns. It is clear that our new approach collects a much more comprehensive set of terms and documents during sampling, and that different collections require samples of greatly varying size.

UDC39. Similar experiments using the UDC39 testbed are shown in Table 6. The same query set is used for experiments on this testbed. Table 6 confirms that our new method outperforms the traditional query based sampling approach; furthermore, our approach is more effective than a central index in many cases. Central index performance has often been viewed as an ideal goal in previous

[3] Some of the collections in this testbed have very few documents(less than 20). We did not use query probing for those collections and consider the whole collection as its summary in both methods.

Table 7. Effectiveness of adaptive sampling on SYM236 with $\eta = 3$ and $\tau = 1\%$

Cutoff	Relevant Retrieved	MAP	P@10	P@20	R-Precision
1	0512	0.0075	0.1392	0.1052	0.0169
10	3191	0.0365	0.2510	0.2281	0.0837
20	4837	0.0580	0.2816	0.2526	0.1176
50	6947	0.0858	0.2796	0.2643	0.1536
118	7803	0.0938	0.2398	0.2352	0.1606

work [Craswell et al., 2000]. Developing a distributed system that outperforms the central index in all cases is still one of the open questions in distributed information retrieval but has been reported as achieveable [French et al., 1999]. According to these results, the performance of our DIR system was greater than the central index for cutoffs 10, 20, and 30 for precision-oriented metrics. For *cutoff* = 10, for example, the system only searches the top 10 collections for each query. This means that it searches only about a quarter of the collections and documents used by the central index, but shows greater effectiveness. Again, values flagged with (*) and (**) indicate statistical significant using the *t-test*.

Changing η and τ. In the results discussed above, we used values for η and τ obtained from our initial experiments. Decreasing η or increasing τ leads to faster termination of query probing, with less effective summaries. In Table 7, we have decreased the threshold τ to 1% — thus increasing the sample sizes — for SYM236. In most cases, the effectiveness is greater than for the same parameters in Table 4, that uses the old τ and η values. Although the results are better, they are more costly. Table 5 shows that the number of documents sampled with $\eta = 1\%$ is about twice that with $\eta = 2\%$. The results for the UDC39 were also tested and found to be similar (but are not presented here).

5 Conclusions

We have proposed a novel sampling strategy for query probing in distributed information retrieval. In almost all previous work on query probing, the sample size was 300 documents; we have shown that such small samples lead to considerable loss of effectiveness. In contrast to these methods, our system adaptively decides when to stop probing, according to the rate of which new unique terms are received. Our results indicate that once the rate of arrival of new terms has become constant, relatively few new significant terms — those of high impact in retrieval — are observed. We compared our new approach and traditional model for query-based sampling on two different testbeds. We found that collections have different characteristics, and that the sample size will vary between collections. The effectiveness of the new approach was not only significantly better than the fixed-size sampling approach, but also outperformed a central index in some cases. While the use of larger samples leads to greater initial costs, there is a significant benefit in effectiveness for subsequent queries.

References

R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999.

P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871, 2003.

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

J. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, 1995. ACM Press.

J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pages 479–490, Philadelphia, Pennsylvania, 1999. ACM Press.

N. Craswell, P. Bailey, and D. Hawking. Server selection on the world wide web. In *Proceedings of the fifth ACM Conference on Digital Libraries*, pages 37–46, San Antonio, Texas, 2000. ACM Press.

D. D'Souza, J. Thom, and J. Zobel. Collection selection for managed distributed document databases. *Inf. Process. Manage.*, 40(3):527–546, 2004a.

D. D'Souza, J. Zobel, and J. Thom. Is CORI effective for collection selection? an exploration of parameters, queries, and data. In P. Bruza, A. Moffat, and A. Turpin, editors, *Proceedings of the Australian Document Computing Symposium*, pages 41–46, Melbourne, Australia, 2004b.

J. French, A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey, and Y. Mou. Comparing the performance of database selection algorithms. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 238–245, Berkeley, California, 1999. ACM Press.

L. Gravano, H. Garcia-Molina, and A. Tomasic. GlOSS: text-source discovery over the internet. *ACM Trans. Database Syst.*, 24(2):229–264, 1999.

L. Gravano, P. G. Ipeirotis, and M. Sahami. Qprober: A system for automatic classification of hidden-web databases. *ACM Trans. Inf. Syst.*, 21(1):1–41, 2003.

P. Ipeirotis. *Classifying and Searching Hidden-Web Text Databases*. PhD thesis, Columbia University, USA, 2004.

P. G. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pages 767–778, Paris, France, 2004. ACM Press.

B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, 2000.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

W. Meng, C. Yu, and K. Liu. Building efficient and effective metasearch engines. *ACM Comput. Surv.*, 34(1):48–89, 2002.

A. L. Powell and J. French. Comparing the performance of collection selection algorithms. *ACM Trans. Inf. Syst.*, 21(4):412–456, 2003.

M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, 2005. ACM Press.

L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–305, Toronto, Canada, 2003. ACM Press.

H. E. Williams and J. Zobel. Searchable words on the web. *International Journal of Digital Libraries*, 5(2):99–105, 2005.

B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 41–50, Melbourne, Australia, 1997. World Scientific Press.