Technical Report IR-154, Center for Intelligent Information Retrieval, Univeriversity of Massachusetts. 1999.

Query-Based Sampling of Text Databases

Jamie Callan, Margaret Connell, and Aiqun Du

Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, Massachusetts 01003-4610, USA

email: callan@cs.umass.edu

Abstract

The proliferation of searchable text databases on corporate networks and the Internet causes a database selection problem for many people. Algorithms such as GlOSS and *CORI* networks can automatically select which text databases to search for a given information need, but only if given a set of *resource descriptions* that accurately represent the contents of each database. However, the existing techniques for acquiring resource descriptions have significant limitations when used in wide area networks controlled by many parties.

This paper presents *query-based sampling*, a new technique for acquiring accurate resource descriptions. Query-based sampling does not require the cooperation of resource providers nor does it require that resource providers use a particular search engine or representation technique. An extensive set of experimental results demonstrate that accurate resource descriptions are created, that computation and communication costs are reasonable, and that the resource descriptions do in fact enable accurate automatic database selection.

1 Introduction

When many document databases are accessible, the first step of Information Retrieval is deciding where to search. Manual selection can be difficult when there are many databases from which to choose, so researchers have developed automatic *content-based* database selection algorithms. A content-based selection algorithm ranks a set of text databases by how well each database matches or satisfies the given query [6, 2, 14, 5, 15, 3]. Content-based database selection has a number of desirable properties, among them reasonable accuracy, scalability, low computational costs, and ease of use.

Database selection algorithms need information about what each database contains. This information, which we call a *resource description*, is simply *assumed* to be available in most prior research. However, in practice, accurate resource descriptions can be difficult to acquire in environments, such as the Internet, where resources are controlled by many parties with differing interests and capabilities. Our interest in this paper is in studying how accurate resource descriptions can be acquired in multi-party environments.

Recent standardization efforts, such as the proposed STARTS extension to Z39.50 [4], illustrate the problem. STARTS requires every resource provider to provide accurate resource descriptions upon request. We call STARTS a *cooperative protocol*, because it only succeeds when each resource provider:

- is able to provide resource descriptions,
- chooses to provide resource descriptions,
- is able to represent database contents accurately, and

- chooses to represent database contents accurately.

Cooperative protocols are appropriate solutions when all resources are controlled by a single party that can mandate cooperation.

In multi-party environments such as the Internet or large corporate networks, complete cooperation is unlikely. Older database systems may be unable to cooperate, some services will refuse to cooperate because they have no incentive or are allied with competitors, and some services may misrepresent their contents, for example, to lure people to the site. All of these characteristics can be found today on the Internet; some of them occur in large corporate networks, too.

One of the most serious problems with cooperative techniques such as STARTS is the great variety in how resource descriptions are created. Most of the prior research is based on descriptions consisting of term lists and term frequency or term weight information [6, 2, 5, 14]. However, differences in tokenizing, case conversion, stopword lists, stemming algorithms, proper name handling, and concept recognition are common, making it impossible to compare term frequency information produced by different parties, even if all parties are able and willing to cooperate.

Resource selection algorithms require accurate resource descriptions. However, the weaknesses of cooperative protocols make them an unsuitable solution for environments where resources are controlled by many parties. In these environments, a more robust solution is required.

This paper presents a new method of acquiring resource descriptions that requires no explicit cooperation from resource providers. Instead, resource descriptions are created as a result of running queries and examining the documents that are returned. This new method, which we call *query-based sampling*, is effective, is efficient, is robust, and can be applied in environments where it is not practical to rely on cooperation.

The next section describes query-based sampling. Sections 3 and 4 describe an extensive set of experiments. Section 5 discusses other uses for query-based sampling, and Section 6 concludes.

2 Query-Based Sampling

Our goal is a method of acquiring resource descriptions that is not overly complex, that does not require special cooperation from resource providers, that can be applied to older ("legacy") systems, that is difficult to deceive, and that is not sensitive to indexing choices made by resource providers.

It is well-known that the characteristics of a population can be determined to a desired degree of accuracy by random sampling. If it were possible to sample documents randomly from a resource, Zipf's Law suggests that the important vocabulary would be discovered fairly rapidly [16]. Random selection requires that each resource provider cooperate, because *the provider* must select documents randomly from its database, so random selection is not a solution. However, it suggests a solution.

The selection service can obtain *biased samples* of each database by running queries and examining the documents returned in response. We call this *query-based sampling*, to emphasize the biased nature of each sample. Query-based sampling satisfies all of the criteria outlined above, because it assumes only that database providers perform their usual service of running queries and returning documents.

Our central hypothesis is that a sufficiently unbiased sample of documents can be constructed from the union of biased samples obtained by query-based sampling.

Query-based sampling is implemented with a simple algorithm, outlined below.

1. Select an initial query term.
2. Run a one-term query on the database.
3. Retrieve the top N documents returned by the database.
4. Update the resource description based on the characteristics of the retrieved documents.
5. If a stopping criterion has not yet been reached,
   (a) Select a new query term; and
   (b) Go to Step 2.

The algorithm involves several specific choices, for example how query terms are selected, how many documents to examine per query, and when to stop sampling. Discussion of these choices is deferred to later sections of the paper.

Table 1: Test corpora.

Name	*Size, in bytes*	*Size, in documents*	*Size, in unique terms*	*Size, in total terms*	*Variety*
CACM	2MB	3,204	6,468	117,473	homogeneous
WSJ88	104MB	39,904	122,807	9,723,528	heterogeneous
TREC-123	3.2GB	1,078,166	1,134,099	274,198,901	very heterogenenous

How best to represent a large document database is an open problem. However, much of the prior research is based on simple resource descriptions consisting of term lists, term frequency or term weight information, and information about the number of documents [6, 5, 14] or number of words [2, 15] contained in the resource. Zipf's Law suggests that the first two pieces of information, term lists and the relative frequency of each term, can be acquired by sampling [16, 10].

It is not clear whether the size of a resource can be estimated with query-based sampling, but it is also not clear that this information is actually required for accurate database selection. We return to this point later in the paper.

The hypothesis motivating our work is that sufficiently accurate resource descriptions can be learned by sampling a text database with simple 'free-text' queries. This hypothesis can be tested in two ways:

1. by comparing resource descriptions learned by sampling known databases (*'learned resource descriptions'*) with the *actual resource descriptions* for those databases, and

2. by comparing resource selection accuracy using learned resource descriptions with resource selection using actual resource descriptions.

Both types of experiments were conducted and are discussed below.

3 Experimental Results: Description Accuracy

The first set of experiments investigated the accuracy of learned resource descriptions as a function of the number of documents examined. The experimental method was based on comparing learned resource descriptions for known databases with the actual resource descriptions for those databases.

The goals of the experiments were to determine whether query-based sampling learns accurate resource descriptions, and if so, what combination of parameters produce the fastest or most accurate learning. A secondary goal was to study the sensitivity of query-based sampling to parameter settings.

The following sections describe the data, the type of resource description used, the metrics, parameter settings, and finally, experimental results.

3.1 Data

Three full-text databases were used:

CACM: a small, homogeneous set of titles and abstracts of scientific articles from the *Communications of the ACM*;

WSJ88: the 1988 *Wall Street Journal*, a medium-sized corpus of American newspaper articles; and

TREC-123: a large, heterogeneous database consisting of TREC CDs 1, 2, and 3, which contains newspaper articles, magazine articles, scientific abstracts, and government documents [9].

These are standard test corpora used by many researchers. Their characteristics are summarized in Table 1.

3.2 Resource Descriptions

Experiments were conducted on resource descriptions consisting of index terms (usually words) and their document frequencies, *df* (the number of documents containing each term).

Stopwords were not discarded when resource descriptions were constructed. However, during controlled testing, learned and actual resource descriptions were compared only on words that appeared in both resource descriptions, which effectively discarded from the learned resource description any word that was considered a stopword by the database. The databases each used the default stopword list of a well-known IR system, which contained 418 very frequent and/or closed-class words.

Suffixes were not removed from words ('stemming') when resource descriptions were constructed. However, during controlled testing, suffixes were removed prior to comparison to the actual resource description, because the actual resource descriptions (the database indexes) were stemmed.

3.3 Metrics

Resource descriptions consisted of two types of information: a *vocabulary*, and *frequency information* for each vocabulary term. The correspondence between the learned and actual vocabularies was measured with a metric called *ctf ratio*. The correspondence between the learned and actual frequency information was measured with the *Spearman Rank Correlation Coefficient*. Each metric is described below.

3.3.1 *Ctf* Ratio

The terms in a learned resource description are necessarily a subset of the terms in the actual description. One could measure how many of the database terms are found during learning, but such a metric is skewed by the many terms occurring just once or twice in a collection [16]. We desired a metric that gave more emphasis to the frequent and moderately-frequent terms, which we believe convey the most information about the contents of a database.

Ctf ratio is the proportion of term occurrences in the database that are covered by terms in the learned resource description. For a learned vocabulary V' and an actual vocabulary V, *ctf* ratio is:

$$\frac{\sum_{i \in V'} ctf_i}{\sum_{i \in V} ctf_i}$$

where ctf_i is the number of times term i occurs in the database (collection term frequency, or *ctf*). A *ctf* ratio of 80% means that the learned resource description contains the terms that account for 80% of the term occurrences in the database.

Note that the *ctf* ratios reported in this paper are not artificially inflated by finding stopwords, because *ctf* ratio was always computed *after* stopwords were removed.

3.3.2 Spearman Rank Correlation Coefficient

The second component of a resource description is document frequency information (*df*), which indicates the relative importance of each term in describing the database. The learned and actual values for *df* probably should not be compared directly, because they are based on examining different numbers of documents. For example, if the true proportion of documents containing a term is 86%, the most accurate estimate possible after seeing 10 documents is 90%, hence a certain amount of error would be built into the metric, and it would vary based on the number of documents sampled.

A more accurate alternative is to rank terms by their frequency of occurrence and then compare the rankings of terms that occur in both the database and the learned resource description. Zipf's Law indicates that there is a predictable relationship between a term's rank and its frequency in the database [16, 10]. Given a term's rank, its frequency can be estimated relatively accurately, and vice versa.

The Spearman Rank Correlation Coefficient is an accepted metric for comparing two rankings [13]. The Spearman Rank Correlation Coefficient is defined as:

$$R = 1 - \frac{6}{n^3 - n} \Sigma d_i^2$$

where d_i is the rank difference of common term i, and and n is the number of terms. Two rankings are identical when the rank correlation coefficient is 1. They are uncorrelated when the coefficient is 0, and they are in reverse order when the coefficient is -1.

Database selection does not require a rank correlation coefficient of 1.0. It is sufficient for the learned resource description to represent the relative importance of index terms in each database to some degree of accuracy. For example, it might be sufficient to know the ranking of a term $\pm 5\%$. Although most database selection algorithms are likely to be insensitive to small ranking errors, it is an open question how much error a given algorithm can tolerate before selection accuracy deteriorates.

3.4 Parameters

Experiments with query-biased sampling require making choices about how query terms are selected and how many documents are examined per query.

In our experiments, the first query run on a database was determined by selecting a term randomly from the TREC-123 vocabulary. The initial query could be selected using other criteria, for example selecting a very frequent term. Several informal experiments found that the choice of the initial query term had no effect on the quality of the resource description learned or the speed of learning, as long as it retrieved at least one document.

Subsequent query terms were chosen by a variety of methods, as described in the following sections. However, in all cases the terms chosen were subject to requirements similar to those placed on index terms in many text retrieval systems: A term selected as a query term could not be a number, and was required to be 3 or more characters long.

We had no hypotheses to guide the decision about how many documents to sample per database query. Instead, a series of experiments was conducted to determine the effect of varying this parameter.

The CACM and WSJ88 experiments presented in this paper were ended after examining 300 documents. The TREC-123 experiments presented in this paper were ended at 500 documents. These stopping criteria were chosen empirically after running several initial experiments, and were biased by our interest in learning resource descriptions from small (ideally, constant) sized samples. Several experiments with each database were continued until several thousand documents were sampled, to ensure that nothing unusual happened.

3.5 Results

Three sets of experiments were conducted to study the accuracy of resource descriptions learned under a variety of conditions. The first set of experiments was an initial investigation of query-based sampling with the parameter settings discussed above. We call these the *baseline experiments*. A second set of experiments studied the effect of varying the number of documents examined per query. A third set of experiments studied the effect of varying the way query terms were selected. Each set of experiments is discussed separately below.

3.5.1 Results of Baseline Experiments

The *baseline experiments* were an initial investigation of query-based sampling. The goal of the baseline experiments was to determine whether query-based sampling produced accurate resource descriptions, and if so, how accuracy varied as a function of the number of documents examined.

The initial query term was selected randomly from the TREC-123 resource description, as described above. Subsequent query terms were selected randomly from the resource description being learned.

The top four documents retrieved by each query were examined to update the resource description. Duplicate documents, that is, documents that had been retrieved previously by another query, were discarded, hence some queries produced fewer than four documents.

Figure 1a shows that query-based sampling quickly finds the terms that account for 80% of the non-stopword term occurrences in each collection.[1] After about 250 documents, the new vocabulary being discovered is terms that are relatively rare. This result is consistent with Zipf's law [16].

[1] Recall that stopwords were excluded from the comparison. If stopwords were included in the comparison, the rate of convergence would be even faster.

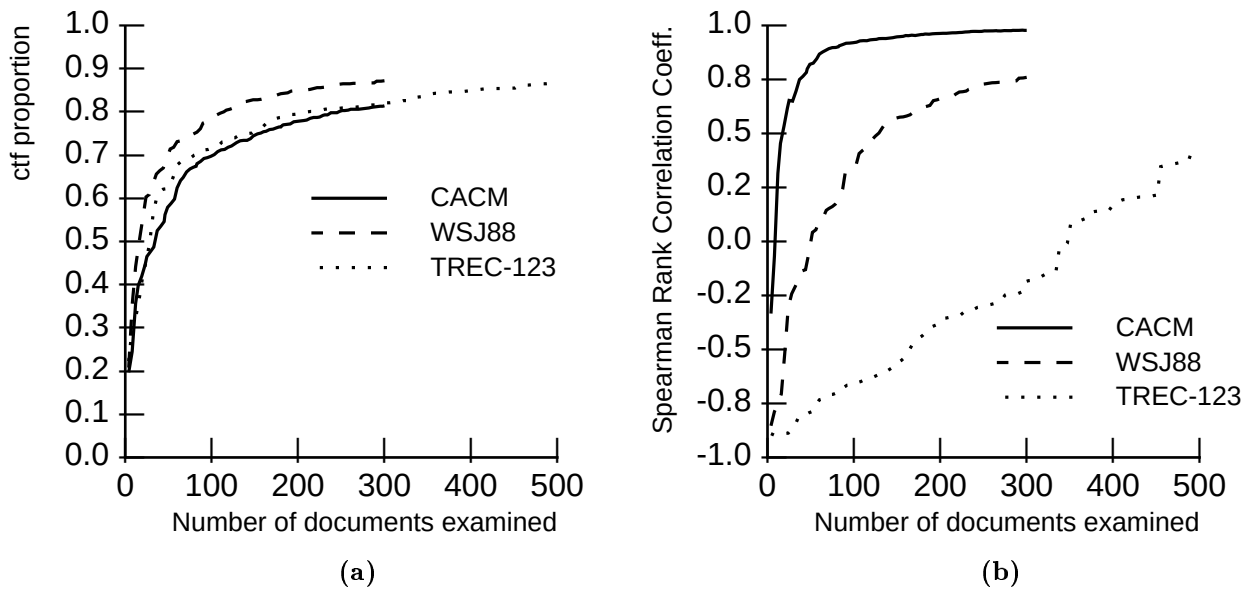

Figure 1: Measures of how well a learned resource description matches the actual resource description of a full-text database. **(a)** Percentage of database word occurrences covered by terms in the learned resource description. **(b)** Spearman rank correlation coefficient between the term rankings in the learned resource description and the database. (Four documents examined per query.)

Table 2: Effect of varying the number of documents examined per query on how long it takes a sampling method to reach a *ctf* proportion of 80%.

Database	Metric	1 Doc/Qry	2 Docs/Qry	4 Docs/Qry	6 Docs/Qry
CACM	ctf	267 docs	251 docs	248 docs	231 docs
WSJ88	ctf	123 docs	123 docs	114 docs	135 docs
TREC-123	ctf	193 docs	185 docs	211 docs	288 docs
CACM	Spearman	.97	.97	.97	.97
WSJ88	Spearman	.40	.43	.43	.47
TREC-123	Spearman	-.27	-.23	-.35	-.43

Figure 1b shows the degree of agreement between the term rankings in the learned and actual resource descriptions, as measured by the Spearman Rank Correlation Coefficient. The degree of correlation achieved at a given number of documents appears to be related to collection size. The smallest collection (CACM) becomes highly correlated quite quickly, while the largest collection (TREC-123) takes significantly longer.

Results from both metrics support the hypothesis that accurate resource descriptions can be learned by examining only a small fraction of the collection. This result is encouraging, because it suggests that query-based sampling is a viable method of learning accurate resource descriptions.

It is interesting that results from the *ctf* ratio suggest that a constant number of documents suffices, whereas results from the Spearman Rank Correlation Coefficient suggest that the number of documents required is partially related to collection size. The reasons for this disagreement are not yet clear. It is possible that vocabulary coverage and frequency information simply converge at different rates.

3.5.2 Results of Varying Sample Size

The baseline experiments sampled the four most highly ranked documents retrieved for each query. However, the sampling process could have retrieved more documents, or fewer documents, per query. Doing so could

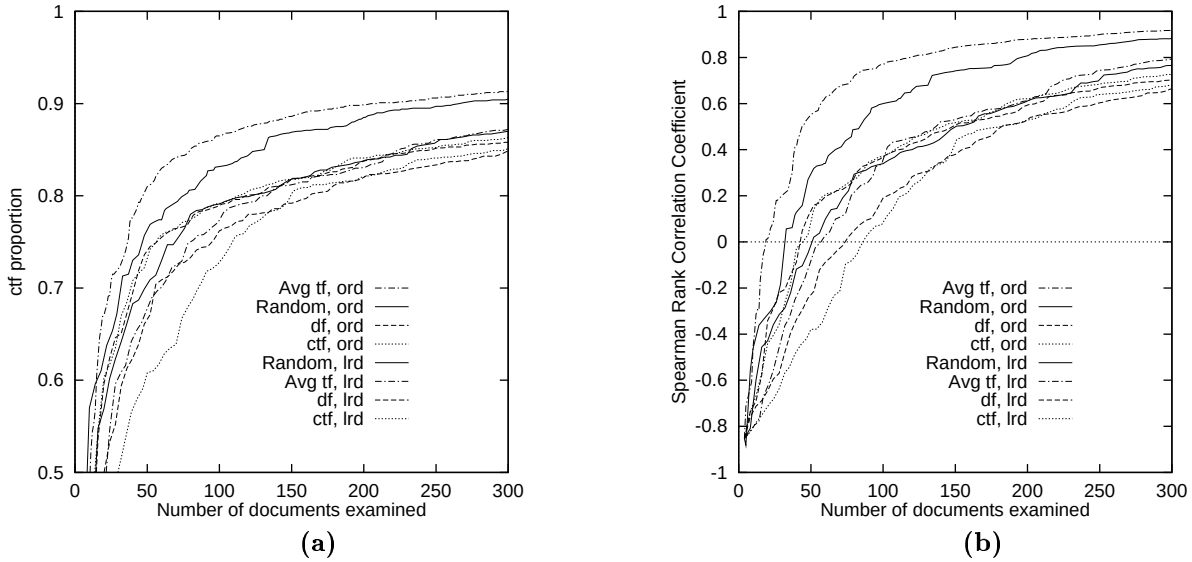

Figure 2: Measures of how different query selection strategies affect the accuracy of a learned resource description. **(a)** Percentage of database word occurrences covered by terms in the learned resource description. **(b)** Spearman rank correlation coefficient between the term rankings in the learned resource description and the database. (1988 Wall Street Journal database. Four documents examined per query.)

change the number of queries and/or documents required to achieve a given level of accuracy, which in turn could affect the costs of running the algorithm.

A series of experiments was conducted to investigate the effects of varying the number of documents examined per query. Values of 1, 2, 4, and 6 were tested.

Graphs summarizing these experiments are nearly indistinguishable, visually, from the graphs in Figure 1, and are therefore not included in this paper. Table 2 shows the number of documents required to reach a *ctf* ratio of 80%. Varying the number of documents examined per query from 1 to 6 causes only minor variations in performance for 2 out of the 3 databases.

In this experiment, larger samples worked well with the small homogeneous collection, and smaller samples worked well with the large heterogeneous collection. We do not find this result surprising. Samples are biased by the queries that draw them; the documents within a sample are necessarily similar to some extent. We would expect that many small samples would better approximate a random sample than fewer large samples in collections where there is significant heterogeneity. The results support this intuition.

3.5.3 Results of Varying Query Selection Strategies

The baseline experiments select query terms randomly from the resource description being learned. Other selection criteria could be used, or terms could be selected from other sources.

One hypothesis was that it would be best to select terms that appear to occur frequently in the collection, i.e., near stopwords, because they would return the most random sample of documents. We tested this hypothesis by selecting frequent query terms, as measured by document frequency (*df*), collection term frequency (*ctf*), and average term frequency (*avg_tf* = *ctf* / *df*).

One early concern was that learned resource descriptions would be strongly biased by the set of documents that just happened to be examined first, and that this bias would be reinforced by selecting additional query terms from the learned resource description. A solution would be to select terms from a different, more complete resource description. This hypothesis was named the *other resource description*, or *ord* hypothesis, and was compared to the default *learned resource description* or *lrd* approach used in the other experiments. The complete TREC-123 resource description served as the 'other' resource description.

A series of experiments was conducted, following the same experimental methodology used in previous

Table 3: The number of queries required to retrieve 300 documents using different query selection criteria.

	Random, ord	Random, lrd	avg_tf, ord	avg_tf, lrd	df, lrd	ctf, lrd
Number of queries	167	78	6,673	107	155	153

experiments, except in how query terms were selected. Query terms were selected either randomly or based on one of the frequency criteria, from either the learned resource description (*lrd*) or the 'other' resource description (*ord*). Four documents were examined per query. Experiments were conducted on all three collections, but results were sufficiently similar that only results for the WSJ88 collection are presented here.

In all of the experiments, selecting terms from the 'other' resource description produced faster learning, as measured by the number of documents examined (Figure 2). However, they also required more queries, sometimes *many* more, to retrieve a given number of documents (Table 3). The difference in the number of queries was due to selecting terms that were either stopwords or that did not occur in the 'other' resource description. Recall also that the 'other' language model was an exact match to one sampled database (TREC-123) and a superset of another (WSJ88). The number of failed queries might have been higher if the 'other' resource description had been a less similar database.

The experiments demonstrate that selecting query terms from the learned resource description, as opposed to a more complete 'other' resource description, does *not* produce a skewed sample of documents. The rate of learning is faster if measured by the number of queries, and slower if measured by the number of documents. Whichever metric is used, a relatively unbiased language model is learned with moderate cost.

The experiments also demonstrate that selecting query terms randomly from the learned resource description is more effective than selecting them based on high frequency. This result was a surprise, because our hypothesis was that high frequency terms would either occur in many contexts, or would have relatively weak contexts, producing a more random sample. This hypothesis was not validated by the experiments.

4 Experimental Results: Selection Accuracy

The experiments described in the previous section investigate how quickly the learned resource description for a database converges upon the actual resource description. However, we do not know how accurate a resource description needs to be for accurate resource selection. Indeed, we do not even know that description accuracy is correlated with selection accuracy, although we presume that it is.

The second set of experiments investigated the accuracy of resource selection as a function of the number of documents examined. The experimental method was based on comparing the effectiveness of the *database ranking algorithm* when using complete and learned resource descriptions. Databases were ranked with the Inquery IR system's default database ranking algorithm [2].

The following sections describe the data, the type of resource description used, the metrics, parameter settings, and finally, experimental results.

4.1 Data

The TREC-123 database described above (Section 3.1) was divided into 100 smaller databases of roughly equal size (about 30 megabytes each). Each database contained documents from a single source, ordered as they were found on the TREC CDs; hence documents in a database were also usually from similar timeframes. CD 1 contributed 37 databases, CD 2 contributed 27 databases, and CD 3 contributed 36 databases.

Queries were based on TREC topics 51-150 [8]. We used query sets INQ001 and INQ026, both created by the UMass CIIR as part of its participation in TREC-2 and Tipster 24 month evaluations [1]. Queries in these query sets are long, complex, and have undergone automatic query expansion.

The relevance assessments were the standard TREC relevance assessments supplied by the U.S. National Institute for Standards and Technology [8].

4.2 Resource Descriptions

Each experiment used 100 resource descriptions (one per database). Each resource description consisted of a list of terms and their document frequencies (*df*), as in previous experiments. Terms on a stopword list of 418 common or closed-class words were discarded. The remaining terms were stemmed with KStem [11].

4.3 Metrics

Several methods have been proposed for evaluating resource selection algorithms [7, 5, 2, 12, 3]. The most appropriate for our needs is a metric sometimes called R [12] or $\hat{R}$ [3] that measures the percentage of relevant documents contained in the n top-ranked databases.

4.4 Parameter Settings

The experiments in Section 3 suggested that any relatively small sample size is effective, and that different choices produce only small variations in results. We therefore chose a sample size of four (4 documents per query), to be consistent with the baseline results in previous experiments.

It was unclear from the experiments in Section 3 when enough samples had been taken. One metric (*ctf* ratio) suggested that about 300 documents were sufficient to build an accurate resource description, while another metric (Spearman Rank Correlation) suggested that the number depended upon the size of the database. We chose to build resource descriptions from samples of 100 documents (about 25 queries), 300 documents (about 75 queries), and 700 documents (about 175 queries) from each database, in order to cover the space of "reasonable" numbers of samples. If results varied dramatically, we were prepared to conduct additional experiments.

The collection ranking algorithm itself forces us to set one additional parameter. The collection ranking algorithm normalizes term frequency statistics ($df_{i,j}$) using the length, in words, of the collection (cw_j) [2]. However, we do not know how to estimate collection size with query-based sampling. In our experiments, term frequency information (*df*) was normalized using the length, in words, of the set of documents used to construct the resource description.

4.5 Experimental Results

The experimental results are summarized in the two graphs in Figure 3 (one per query set). The baseline in each graph is the curve showing results with the actual resource description ("complete resource descriptions"). This is the best result that the collection ranking algorithm can produce when given a complete description for each collection.

Our interest is in the difference between what is achieved with complete information and what is achieved with incomplete information. Both graphs show only a small loss of effectiveness when resource descriptions are based on 700 documents. Losses grow as less information is used, but the loss is small compared to the information reduction. Accuracy at "low recall", i.e., when only 10-20% of the databases are searched, is quite good, even when resource descriptions are based on only 100 documents.

These results are consistent with the results presented in Section 3. The earlier experiments showed that term rankings in the learned and actual resource descriptions were highly correlated on the WSJ88 database after examining 100-300 documents. Each database in this experiment was about one third the size of the WSJ88 database, hence the learned and actual term rankings were very highly correlated.

These experimental results also demonstrate that it is possible to rank collections without knowing their sizes. The decision to replace information about collection size with information based on the sizes of sampled documents appeared effective. A more thorough test of this decision would be collection selection over a set of resource descriptions built from different numbers of documents. However, the sampling process can choose whether to examine a constant or varying number of documents per resource, so a more thorough test may be of only academic interest.

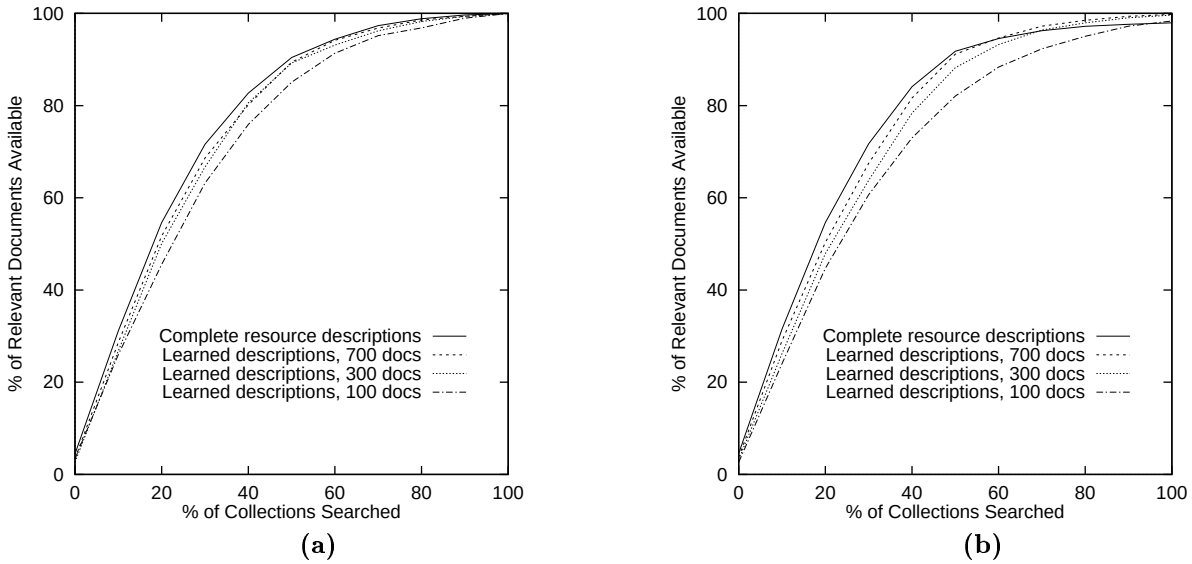

Figure 3: Measures of collection selection accuracy using resource descriptions of varying accuracy. **(a)** Topics 51-100 (TREC query set INQ026). **(b)** Topics 101-150 (TREC query set INQ001). (4 documents examined per query. TREC volumes 1, 2, and 3.)

5 Other Uses

The set of documents sampled from a single database reflects the contents of that database. One use of these documents is to build a resource description for a single database, as described above. However, other uses are possible.

One potential use is in a query expansion database. Recent research shows that query expansion significantly improves the accuracy of database selection [15]. The state-of-the-art in query expansion is based upon analyzing the searched corpus for cooccurrence patterns, but what database(s) should be used when the task is database selection? This question has been unanswered.

If the documents sampled from each database were combined into a query expansion corpus, the result would be a set of documents that reflects the contents and word cooccurrence patterns across *all* of the available databases. It would require little additional effort for a database selection service to create a query expansion database in this manner.

Cooccurrence-based query expansion can be viewed as a form of data mining. Other forms of data mining could also be applied to the set of documents sampled from all databases. For example, frequent concepts, names, or relationships might be extracted and used in a visualization interface.

The ability to construct a single database that acts as a surrogate for a set of databases is significant, because it could be a way of rapidly porting many familiar Information Retrieval tools to environments containing many databases. Although there are many unanswered questions, this appears to be a promising direction for future research.

6 Conclusions

Our hypothesis was that an accurate description of a text database can be constructed from documents obtained by running queries on the database. The experiments presented in this paper confirm that hypothesis. The resource descriptions created by *query-based sampling* are sufficiently similar to resource descriptions created from complete information that it makes little difference which is used for database selection.

Query-based sampling avoids many of the limitations of cooperative protocols such as STARTS. Query-based sampling can be applied to older ('legacy') databases and to databases that have no incentive to cooperate. It is not as easily defeated by intentional misrepresentation. It also avoids the problem of needing

to reconcile the differing tokenizing, stopword lists, word stemming, case conversion, name recognition, and other representational choices made in each database. These representation problem are perhaps the most serious weakness of cooperative protocols, because they exist even when all parties *intend* to cooperate.

The experimental results also demonstrate that the cost of query-based sampling, as measured by the number of queries and documents required, is reasonably low, and that query-based sampling is robust with respect to variations in parameter settings.

Several open questions remain, among them whether the number of documents in a database can be estimated with query-based sampling. We have shown that this information is not required for database selection, but it is nonetheless desirable information. It is also an open question how many documents must be sampled from a resource to obtain a description of a desired accuracy, although 300-500 documents appears to be very effective across a range of database sizes.

The work reported here can be extended in several directions, to provide a more complete environment for searching and browsing among many databases. For example, the documents obtained by query-based sampling could be used to provide query expansion for database selection, or to drive a summarization or visualization interface showing the range of information available in a multi-database environment. More generally, the ability to construct a single database that acts as a surrogate for a large set of databases offers many possibilities for interesting research.

Acknowledgements

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement EEC-9209623, and by the U.S. Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract F19628-95-C-0235. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor(s).

References

[1] J. P. Callan, W. B. Croft, and J. Broglio. TREC and TIPSTER experiments with INQUERY. *Information Processing and Management*, 31(3):327–343, 1995.

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

[3] J.C. French, J.C. Powell, C.L. Viles, T. Emmitt, and K.J. Prey. Evaluating database selection techniques: A testbed and experiment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1998.

[4] L. Gravano, K. Change, H. García-Molina, and A. Paepcke. STARTS Stanford proposal for Internet meta-searching. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, 1997.

[5] L. Gravano and H. García-Molina. Generalizing GLOSS to vector-space databases and broker hierarchies. In *Proceedings of the 21st International Conference on Very Large Databases (VLDB)*, pages 78–89, 1995.

[6] L. Gravano, H. García-Molina, and A. Tomasic. The effectiveness of GLOSS for the text database discovery problem. In *Proceedings of SIGMOD 94*, pages 126–137. ACM, September 1994.

[7] L. Gravano, H. García-Molina, and A. Tomasic. Precision and recall of GLOSS estimators for database discovery. Technical Report STAN-CS-TN-94-10, Computer Science Department, Stanford University, 1994.

[8] D. Harman, editor. *The Second Text REtrieval Conference (TREC2).* National Institute of Standards and Technology Special Publication 500-215, Gaithersburg, MD, 1994.

[9] D. Harman, editor. *Proceedings of the Third Text REtrieval Conference (TREC-3).* National Institute of Standards and Technology Special Publication 500-225, Gaithersburg, MD, 1995.

[10] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects.* Academic Press, New York, 1978.

[11] R. Krovetz. *"Word Sense Disambiguation for Large Text Databases.* PhD thesis, University of Massachusetts at Amherst, 1995.

[12] Z. Lu, J.P. Callan, and W.B. Croft. Measures in collection ranking evaluation. Technical Report 96-39, Department of Computer Science, University of Massachusetts, 1996.

[13] M.J. Moroney, editor. *Facts from figures.* Penguin, Baltimore, 1951.

[14] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, Gaithersburg, MD, 1995. National Institute of Standards and Technology, Special Publication 500-225.

[15] J. Xu and J. Callan. Effective retrieval of distributed collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, 1998. ACM.

[16] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley, Reading, MA, 1949.