# Determining Stopping Criteria in the Generation of Web-Derived Language Models

Gary A. Monroe David R. Mikesell James C. French*
Department of Computer Science
University of Virginia
Technical Report CS-2000-30

May 10, 2000

**Abstract**

In this work, we present a small-scale evaluation of two query-based sampling techniques for building language models, using a database comprised of world-wide web documents. We propose a metric by which it is possible to determine when to cease sampling a given web database, and we compare this new metric to other metrics that have been used in previous work to determine the fidelity of sampled language models.

## 1 Introduction

With the recent and explosive growth of the World Wide Web, many new repositories of information have become available to the public. With so many databases available for search, one potential problem that people face is how to decide which database would be appropriate to query. When there are only a small number of text databases available, persons can either search all the databases for the desired information or they can simply familiarize themselves with the general contents of each database and direct their query to one or more specific collections. However, when there are hundreds or thousands of searchable databases and the content of those databases cannot be easily ascertained, automatic database selection algorithms can be used to assist in the choice of which databases to search by identifying those databases that are most likely to satisfy the information need [GGM95, CLC95, FPV98, FPC+99b, FPC99a, XC99].

Database selection algorithms require information about the contents of those databases over which they are selecting. A number of ways have been proposed for defining what information is required and how to acquire it, cf. [GCGMP97, HT99, CCD99]. Following Callan et al.[CCD99] we refer to this information as a *language model*. Our language model of a database lists the words that occur in the database and their frequency of occurrence and, perhaps, other information. Since a completely accurate language model must include a representation of all of the words from all of the documents in a database, access to all of the documents in the database is required to generate this base language model. Because many databases do not provide this level of access to their contents (especially web databases), methods have been proposed for acquiring such metadata automatically and without the need for cooperation from the database. One such method is called query-based sampling and is the focus of the research presented in this paper. Query-based sampling is a

*french@cs.virginia.edu

sampling technique in which metadata is inferred by interacting with each database and observing the outcomes.

Some previous research has been done on query-based sampling but it is a relatively new technique and little is known about the generality and behavior of the technique outside a laboratory environment. Prior research by Callan, Connell, and Du [CCD99] demonstrated this technique's effectiveness at learning accurate metadata for several research testbeds of varying size and heterogeneity. These results were promising but the conditions under which they studied query-based sampling may not have been appropriate to generalize the results to all text databases. The research presented in this paper investigates the use of query-based sampling by examining its effectiveness in generating accurate language models of databases composed of World Wide Web data.

## 2 Research Questions

Prior research with laboratory testbeds showed that the query-based sampling technique was able to create language models that were representative of the content of the sampled databases [CCD99]. However, it is not known how well query-based sampling will perform on databases comprised of world-wide web documents. It is also unknown what degree of correlation is required in order to effectively use that information in database selection algorithms, or even how to correctly measure that correlation. In addition, if it is determined that query-based sampling is a viable technique for learning language models of web databases, then, in order for this technique to be efficient, there must be a way to determine when a sampled language model has evolved to the point where it is a "good enough" representation of the underlying data.

The preliminary study presented here is intended to address questions about the utility of query-based sampling to web databases. We examine the evolution of a language model sampled from a database comprised of web pages, and attempt to determine if a learned language model highly correlated to the actual web database language model could be discovered in a reasonable sample size. Moreover, we were interested in examining statistics that might eventually prove useful in establishing a stopping criterion for the sampling procedure.

## 3 Methodology

Three different language model sampling strategies were implemented and evaluated over world-wide web data using a variety of metrics. Models created by the different strategies were examined over their evolution as more documents were sampled by comparison to the complete language model, and by comparison to the model's predecessor in its evolution. By comparing to the complete language model, we hoped to gain insight into the rate of convergence of a given model to the complete model. By comparing a model to its predecessor in its evolution, we hoped to determine a stopping criterion.

Prior work evaluated the fidelity of learned language to the actual language models using the *ctf* ratio, percentage of the total vocabulary learned, and Spearman rank correlation coefficient[CCD99]. We also used these metrics in this work, but since they require full knowledge of the actual language model they cannot be used to determine a stopping criterion in practice. However, they do provide insight as to how the learned model converges to the complete model.

A number of metrics that could be known in a real-world sampling situation were also examined. The growth of the size of the model and the rate of change of the model growth provided insight as to the effectiveness of the sampling techniques in gathering vocabulary. The differences in

document frequency proportions were examined to see how quickly document frequencies stabilized with respect to the preceding language model. Finally, the proportion of rare terms in the learned language model was investigated, and how this proportion varied as the language model grew was studied.

### 3.1 Data

The study was conducted on a local database of web documents. The database was created from data provided by the Open Directory Project[1] (ODP). The Open Directory Project is an attempt to create a comprehensive directory of the web by employing the help of thousands of volunteer editors. The directory is organized similarly to category structures of some of the most popular search engines, and some popular search engines use the ODP information in their own categories. ODP provides a comprehensive list of the URLs that have been assigned to each individual subdirectory.

The database was created from the list of URLs in the ODP Recreation/Autos subdirectory. At the time of the experiments, 6620 URLs were listed. We attempted to fetch all of the listed pages and store them locally to maintain a static database to run our experiments. Out of the 6620 listed URLs, 140 were invalid URLs (either return a dead link or an invalid DNS entry), giving a total of 6480 web pages available.

The actual language model was then built from the database of locally stored pages using LAMB[OF00], a language model builder designed for modelling WWW resources using query-based sampling. Each page was fetched, and pages of less than 100 bytes in size were discarded.[2] For the pages of size greater than 100 bytes, all HTML tags and scripts were parsed out of the pages. The remaining text from each page was stopped and stemmed. Each term that was extracted from each page was added to the vocabulary list and the document frequency was calculated. The resulting actual language model was a compilation of 5153 pages. The vocabulary size was 37,691 terms, 58% of which had a document frequency (also referred to as df) of 1.

### 3.2 Search Engine

Once a local database representing an actual web database was created, a local search engine to index and search the local database was needed. MiniSearch v.0.2, a search engine freely available on the web[3], was used as the local search engine to conduct the experiments. MiniSearch consists of two Perl scripts and a few supporting configuration files. One script performs the indexing on a collection of documents. The second script performs searches on the indexed documents.

The search engine models a Boolean search. Query terms can use the logical OR or AND operators. Documents that contain any of the query terms are returned as potential matches. If the OR operator is used (the default), all the potential matches are members of the result set. If the AND operator is used, only the documents that contain all of the query terms are members of the result set. The result set is then ordered by the sum of the term frequencies of the query terms in the document, with the highest sum presented first. The number of results specified by the user is then displayed. Documents with equivalent term frequency sums are displayed in the order in which they were indexed. During this study, only single term queries were used, so results were ordered by term frequency of the query term.

[1] More information is available at http://www.dmoz.org.

[2] Very short pages are often simply frame pages and don't contain any useful content for language modelling purposes.

[3] More information on MiniSearch is available at http://www.dansteinman.com.

### 3.3 Sampling Strategies

Three different sampling strategies were used in the experiment. Two of the strategies were variations on the query-based sampling technique. The initial query term for the query-based sampling techniques was a term taken at random from the complete language model. The third strategy was a purely random sampling of documents.

#### 3.3.1 Random Query-Based Sampling

After the top 10 results were returned from an initial seed query, the vocabulary from the result pages was accumulated. One term was chosen at random from the learned vocabulary, and that term was used for the subsequent query; previously-selected terms were not reused. The probability of a term being chosen was inversely proportional to the number of terms in the sampled language model at that point in its evolution. The top 10 results retrieved by the search engine were examined. If the search engine returned a page had been previously processed by an earlier query, it was discarded. After approximately every 30 pages (actually, from 30-39), a snapshot of the evolving language model was saved. This process was repeated until the vocabulary from 2000 unique pages had been added to the model.

#### 3.3.2 Random Query-Based Sampling (no df =1 query terms)

This strategy is identical to the Random Query-Based Sampling Strategy for the first approximately 25 pages returned by the search engine. After that point, if the randomly chosen term to be used for the next query has a document frequency of 1, then it is discarded and another term is chosen at random. Snapshots of the evolving model were taken as above, and sampling was continued as above until the vocabulary from 2000 unique pages had been added to the model. The thought behind this technique is that, by removing terms with document frequencies equal to one, a higher fidelity model would be produced more rapidly since we believed that rare query terms would tend to point to pages that are less representative of the underlying database.

#### 3.3.3 Completely Random Document Sampling

This technique can be described as picking pages out of a hat. One page was chosen at random without replacement from the 5153 pages used to build the actual language model. The vocabulary from each page was accumulated into the evolving language model. After exactly every 30 pages seen, a snapshot of the evolving language model was saved. This sampling strategy continued until all 5153 pages were chosen. This strategy models random sampling and provides us with a basis for statistical analysis.

For each sampling strategy, 18 different learned language models were built from 18 separate runs. Statistics derived from the models reflect these 18 trials.

## 4 Analysis

As can be seen in figures 1 and 2, the language models that use the query sampling technique grow quickly early in their evolution, then the growth levels off. However, even after 2000 documents have been examined, the models are still growing at nearly 200 terms per 30 documents. Considering that the complete language model has 37,691 terms, this rate of growth is considerable. Note that the models created by randomized document selection grow at a slower rate. This may be due to a

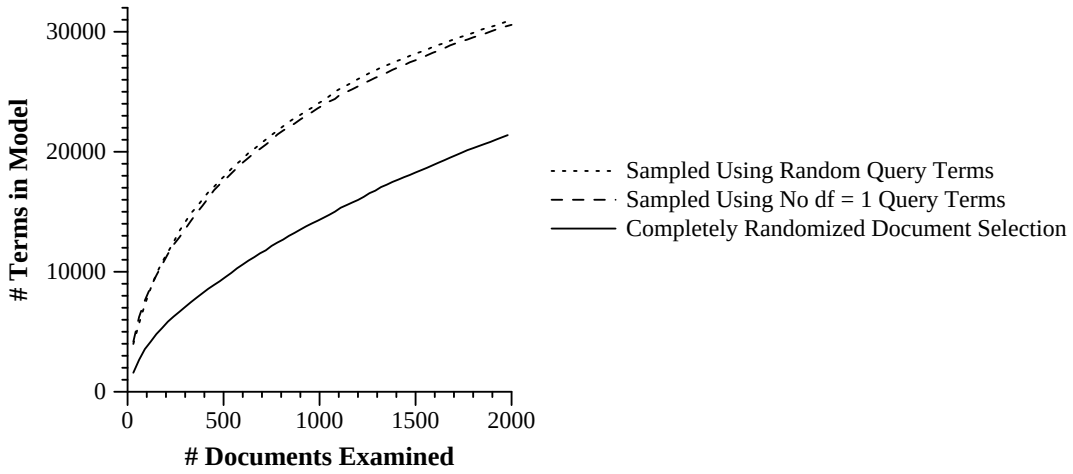


Figure 1: Documents Sampled vs. Number of Terms in Model

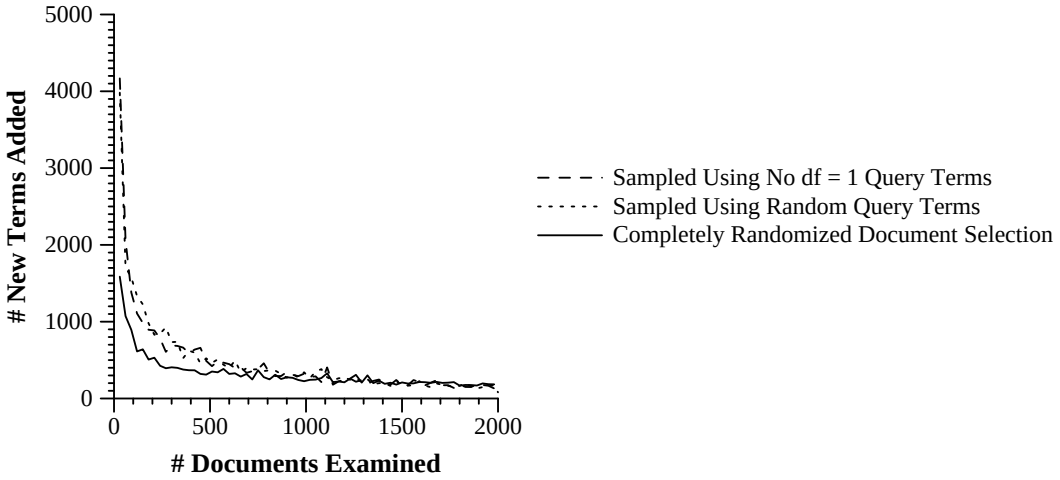


Figure 2: Documents Sampled vs. Number of New Terms Added to Model

size bias introduced by the search engine used by the other techniques. As mentioned before, the search engine sorts according to term frequency, which may be correlated to document size; large documents contain more text and that leads to faster vocabulary growth.

Callan, et al[CCD99], introduced the *ctf* ratio metric as a way to measure the quality of a database by weighting the importance of each term. This ratio is defined as:

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

where $ctf_i$ is the number of times term $i$ occurs in the database; $V'$ denotes the sampled language model and $V$ denotes the complete language model. As figure 3 shows, the *ctf* ratio grows rapidly, and after only 300 documents for the query-based sampling techniques, it approaches 0.9. This indicates that, for this database, the high-frequency terms from the base language model are learned quickly. Unfortunately, to calculate the *ctf* ratio, the collection term frequencies are needed, and that information is not likely to be available in a real-world situation. Note that the sampling technique that eliminates rare query terms produces models with a slightly greater *ctf* ratio than

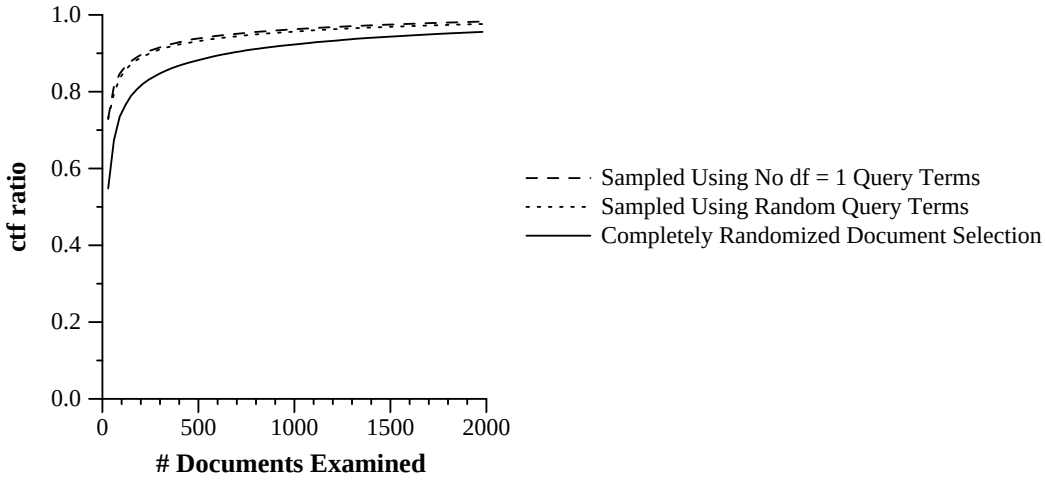


Figure 3: Number of Documents vs. *ctf* Ratio

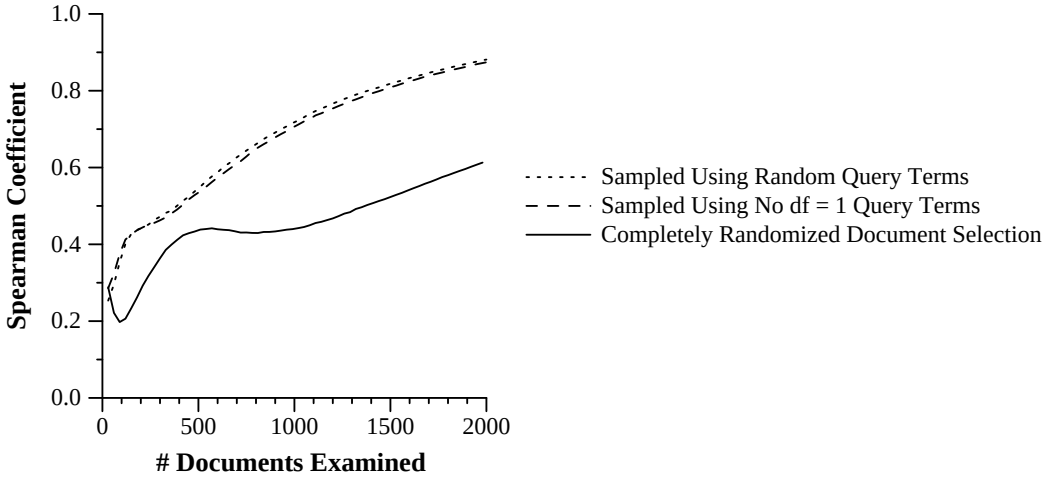


Figure 4: Number of Documents Sampled vs. Spearman Rank Correlation Coefficient

the random query term technique, even though the random query term technique produces slightly larger models for a given number of documents.

Since Zipf's law [Zip49] states that there is a predictable relationship between a term's rank and it's frequency in the database, a comparison of ranks between a sampled model and the complete model may provide some useful information. The Spearman rank correlation coefficient can be used to compare two rankings. However, in this case, its utility is suspect. Since approximately 58% of the terms in the complete model have term frequencies of 1, this indicates that there will be a very large number of ties. Figure 4 shows that, for the query sampled models, the Spearman coefficient rises sharply early, then continues to rise at a slower rate. For the randomized document selection, the Spearman coefficient shows some oscillating behavior until about 1000 documents. Again, this metric has the additional disadvantage of requiring the complete language model in order to calculate it.

The proportion of terms with document frequency of 1 (referred to hereafter as df1 proportion) provides an estimator of how prevalent rare terms are in the language model. This metric does not require knowledge of the complete language model; however, knowledge of the complete model does

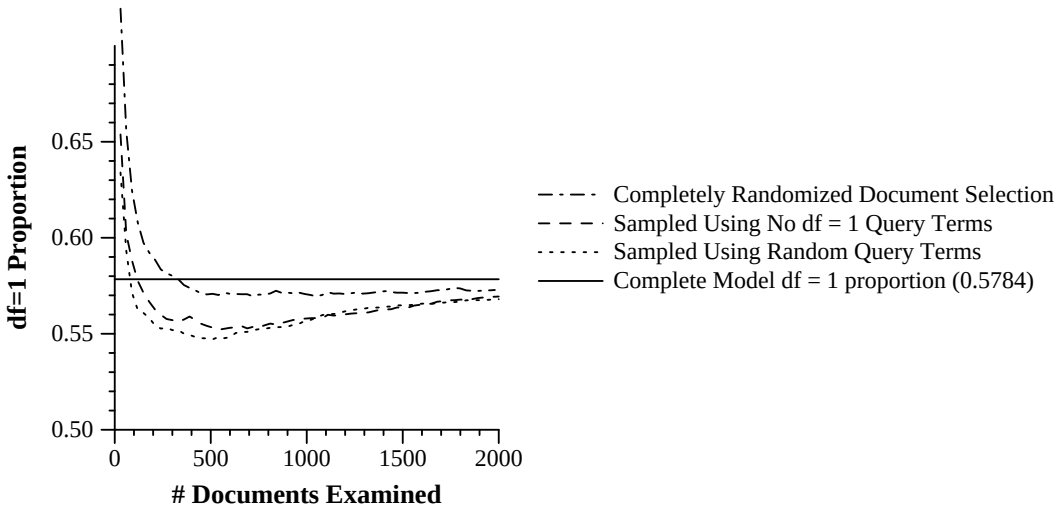


Figure 5: Number of Documents Sampled vs. Proportion of Terms With df = 1

provide information as to what df1 proportion the learned model should converge to. In models from all three sampling techniques, early in the model growth this proportion is overestimated. Then the proportion undershoots the expected value, then slowly approaches the expected value as the models grow. The completely randomized document selection undershoots the expected value the least, which is consistent with a sampling technique that is unbiased. The other two techniques are not random, and exhibit a more dramatic undershoot. All three techniques reach a minimum at approximately 500 documents. After this point, the proportion of rare terms converges slowly to that of the complete model.

Early in the evolution of a model, it is not unreasonable to assume that the model is a poor reflection of the complete model, and is exhibiting some sort of startup behavior until it begins to stabilize. The *ctf* ratio and Spearman coefficient metrics support this assumption, as early on in the growth of a model, these metrics have low values. If we consider the df1 proportion to give an indication of the "poorness" of the model early in its evolution, then it would not be unreasonable to look for a defining point in the df1 proportion graph where we can consider startup behavior to have ended. At the minimum value df1 proportion, we can see from Figures 6 and 7 that the *ctf* ratio is well above 0.85 for all three sampling techniques, and the Spearman rank correlation is between 0.45 and 0.55.

The final metric we examined was the change in RMS difference of df proportions between a language model and its predecessor in its evolution. This is calculated by:

$$\sqrt{\frac{1}{|LM_{j-1}|}\sum_{i \in LM_{j-1}} \left( \frac{df_{i,j}}{n_j} - \frac{df_{i,j-1}}{n_{j-1}} \right)^2}$$

where $LM_j$ represents the language model at a given snapshot, and $LM_{j-1}$ represents the language model at the previous snapshot.

This metric is an indicator of how much a model is changing as it evolves. As expected, the model changes rapidly at the beginning of its evolution, and then levels off quickly. This observation supports the assumption above that models exhibit startup behavior before they stabilize. At the 500 document point, the RMS difference between models is approximately 0.01 for the query-based sampling techniques (see Figure 8).

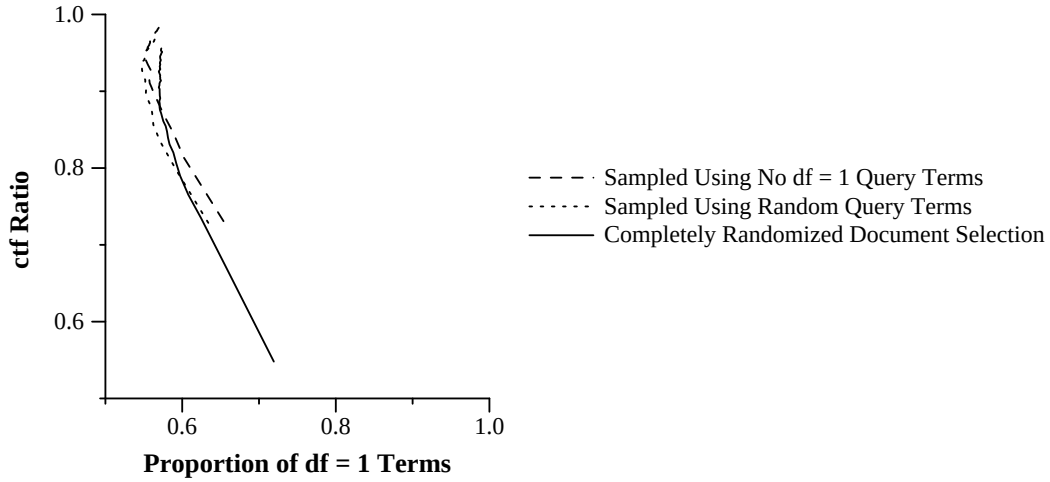


Figure 6: Proportion of Terms With df = 1 vs. *ctf* Ratio

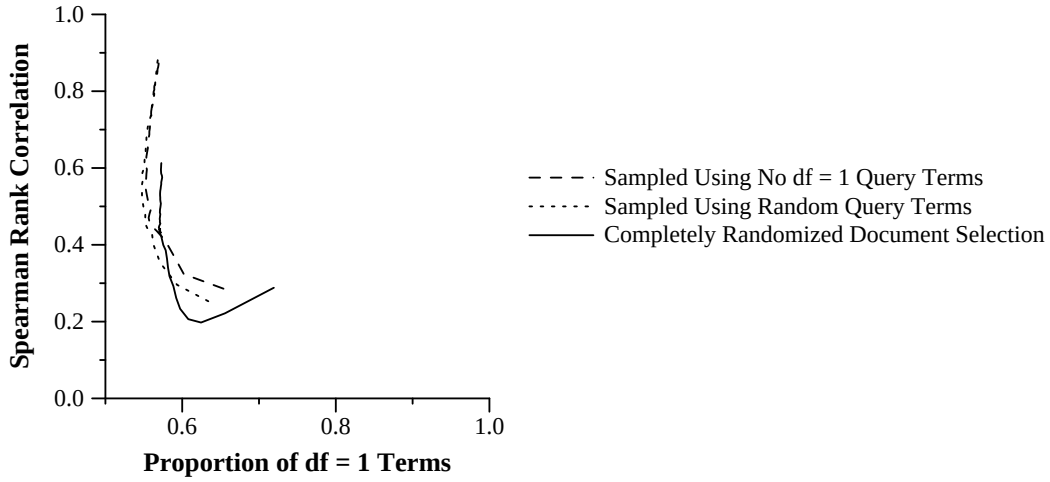


Figure 7: Proportion of Terms with df = 1 vs. Spearman Rank Correlation Coefficient

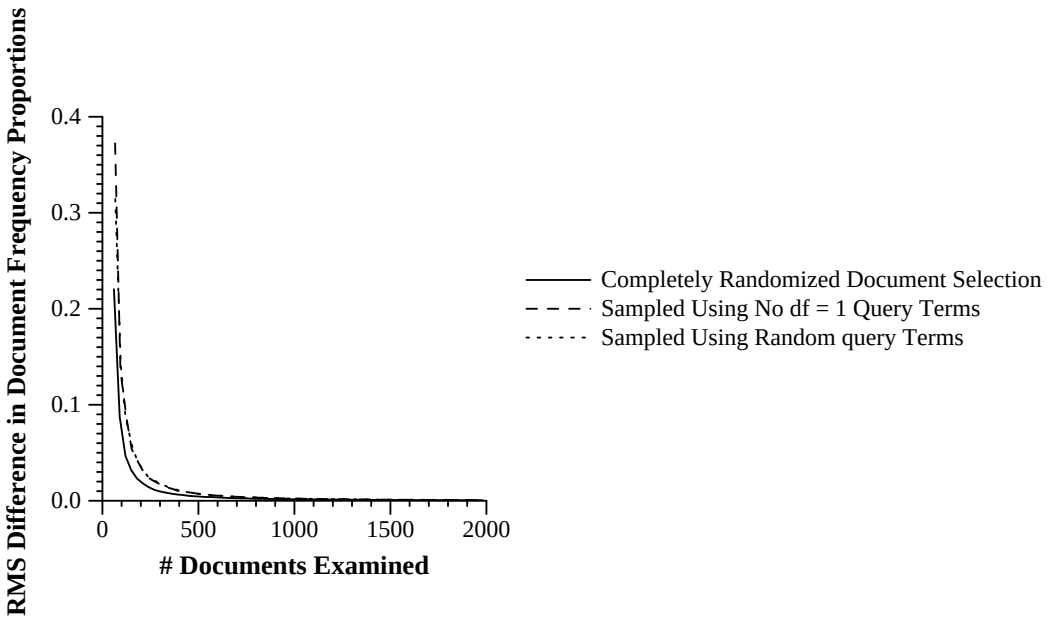


Figure 8: Number of Documents Sampled vs. RMS Difference in Document Frequency Proportions

## 5 Conclusion and Discussion

In the query-based sampling technique to develop learned language models, the question of when to stop querying is troubling. If one stops too soon, then the model is a poor reflection of the underlying data. Stopping too late is wasteful of resources. In this paper, we have examined the application of this technique to web-based data. We have shown that this technique can produce models that reflect the underlying data, and have tested how faithful of a reflection they are. Furthermore, we have presented a metric that may be useful in providing a stopping point for real-world query-based sampling. In other work we have also shown that language models developed by query-based sampling can be used effectively for database selection among web resources[SPM+00].

Eliminating rare terms from the set of query term candidates had little effect on the models. Models generated by eliminating rare query terms were, on the average, very slightly smaller, had slightly higher *ctf* ratios, and, after the startup period, had slightly lower Spearman coefficients than their counterparts generated from a random query term selection. They did have a somewhat higher minimum df1 proportion than models generated from a random query term selection, but whether that difference is useful remains to be seen.

This is a small study, focusing on one very small subset of the web. A future study to investigate these sampling techniques and the utility of the df1 proportion and RMS difference metrics would need to examine a larger subset of the web in order to draw more useful conclusions.

## References

[CCD99] Jamie Callan, Margaret Connell, and Aiqun Du. Automatic discovery of language models for text databases. In *Proceedings of the ACM–SIGMOD International Conference on Management of Data*, pages 479–490, 1999.

[CLC95] J.P. Callan, Z. Lu, and W.B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th International Conference on Research and Development in Information Retrieval*, pages 21–29, 1995.

[FPC99a] J.C. French, A.L. Powell, and J.P. Callan. Effective and efficient automatic database selection. Technical Report CS-99-08, Department of Computer Science, University of Virginia, 1999.

[FPC+99b] J.C. French, A.L. Powell, J.P. Callan, C.L. Viles, T. Emmitt, K.J. Prey, and Y. Mou. Comparing the performance of database selection algorithms. In *Proc. 22nd SIGIR*, pages 238–245, 1999.

[FPV98] J.C. French, A.L. Powell, and C.L. Viles. 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*, pages 121–129, 1998.

[GCGMP97] Luis Gravano, Chen-Chuan K. Chang, Hector Garcia-Molina, and Andreas Paepcke. Starts: Stanford proposal for internet meta-searching. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 207–218, May 1997.

[GGM95] L. Gravano and H. Garcia-Molina. Generalizing gloss to vector-space databases and broker hierarchies. In *Proceedings of the 21st International Conference on Very Large Databases*, 1995.

[HT99] David Hawking and Paul Thistlewaite. Methods for Information Server Selection. *ACM Transactions on Information Systems*, 17(1):40–76, 1999.

[OF00] E. K. O'Neil and J. C. French. A description of the lamb web-derived language model builder. Technical Report CS-2000-31, Department of Computer Science, University of Virginia, May 2000.

[SPM+00] R. Srinivasa, T. Phan, N. Mohanraj, A. L. Powell, and J. C. French. Database selection using document and collection term frequencies. Technical Report CS-2000-32, Department of Computer Science, University of Virginia, May 2000.

[XC99] J. Xu and W.B. Croft. Cluster-based language models for distributed retrieval. In *Proc. 22nd SIGIR*, pages 254–261, 1999.

[Zip49] G. K. Ziph. *Human Behavior an dthe Principle of Least Effort: An Introduction to Human Ecology.* Addison–Wesley, Reading, MA, 1949.