J.T. Yao, V.V. Raghvan, G.Y. Wang (Eds.): WSS'04, pp. 81-87, 2004.

Estimating Size of Search Engines in an Uncooperative Environment

Surendra Karnatapu, Karthik Ramachandran, Zonghuan Wu,
Biren Shah, Vijay V. Raghavan, Ryan Benton
The Center for Advanced Computer Studies
University of Louisiana at Lafayette
{skk0487, kxr3869, zwu, bshah, raghavan, rgb8817}@cacs.louisiana.edu

Abstract

The number of documents that are indexed by a search engine is referred to as the size of the search engine. The information about the size of each underlying search engine is essential for any metasearch engine to conduct search engine selection, result merging and a few other processes. Thus, effectively estimating the size of search engines is important for a metasearch engine that incorporates multiple autonomous search engines. In this paper, we propose an algorithm that achieves better accuracy compared to the other existing methods for estimating the size of search engines, without losing efficiency. Compared to the Sample-Resample approach, which is the best-known approach in literature, our technique also shows much better tolerance to unfavorable environments.

1. Introduction

One of the major observations in distributed information retrieval in the past few years has been that no single search engine indexes the entire web or even a large portion of it [7]. This observation has led to the development of integrated tools to create metasearch engines, which are built on a number of individual search engines. With each of the search engines indexing certain part of the web, a metasearch engine can achieve better coverage by concurrently searching many search engines. But to be able to do so, the builders of the meta-search engine must address two key issues. These are 1) acquiring information about each of the search engines (*Resource Description*), and 2) selecting a subset of the resources (underlying search engines) for a given query (*Resource Selection*). The metasearch engine then merges the ranked results returned by the different search engines before presenting it to the user (*Result Merge*). Statistical approaches have been widely used for addressing the above issues in current metasearch engine systems. One essential piece of information required by such approaches is the size of each of the underlying search engines, i.e. the number of documents indexed by each of the search engines. Usually, it is difficult for a metasearch engine to obtain this information when search engines are not cooperative and hence do not provide the required information. In such situations, it becomes necessary to develop techniques that can estimate the size of search engines.

A few methods have been proposed to estimate the size of a search engine in uncooperative environments and they will be briefly reviewed in section 2. In section 3, we propose an approach for Boolean search engine systems that provides higher estimation accuracy than the best available method with comparable efficiency. In section 4, we explain our experimental setup. In Section 5, we explain our results, analyze them and highlight key reasons as to why our approach scores over the others. Finally, in Section 6, we conclude and present the future work, which is a sketch of our effort to further validate and improve our approach.

2. Related Work

To the best of our knowledge, there are three algorithms that have been proposed for estimating the size of a search engine. They are Interval Estimation based on Sample Data [1], Capture-Recapture [2], and Sample-Resample [3]. In this section, we briefly review the three approaches.

The Interval Estimation based on Sample Data [1] technique uses a pair of independent query terms, say (t_1, t_2), to estimate the size of a search engine. The number of documents containing either of the terms (t_1 or t_2) and the number of documents containing both the terms (t_1 and t_2) are found by sending distinct queries to the search engine. The estimate is then computed using probabilistic independence criterion. However, the problem of finding

independent terms is not trivial and was not discussed in the paper. The author manually created a list of term pairs; the two terms in each pair are assumed to be independent. To achieve even reasonable accuracy, averaging the estimate values over a number of individual estimates becomes necessary. This degrades the efficiency of the technique.

The Capture-Recapture [2] technique assumes there are two (or more) independent samples from a population. Let N be the population size, A be the event that an item is included in the first sample, which is of size n_1, B be the event that an item is included in the second sample, which is of size n_2, and m_2 be the number of items that appears in both samples. The probabilities of events A and B, and the relationship between them, are shown below.

$$\mathrm{P}(A) = \frac{n_1}{\mathrm{N}} \quad (1)$$

$$\mathrm{P}(B) = \frac{n_2}{\mathrm{N}} \quad (2)$$

$$\mathrm{P}(A|B) = \frac{m_2}{n_2} \quad (3)$$

Since the two samples are assumed to be independent,

$$\mathrm{P}(A|B) = \mathrm{P}(A) \quad (4)$$

Thus, the population size is estimated as

$$\mathrm{N}_{est} = \left(\frac{n_1 * n_2}{m_2} \right) \quad (5)$$

The above technique was applied to estimate the size of a search engine by sending random queries to the search engine and then sampling from the result documents (ids). Since this method depends heavily on the number of probe queries and the sample documents to achieve good accuracy, a large number of sample queries need to be sent; hence, the technique might not scale well for large search engines [3]. Also, the technique estimates the search engine size using formula (5), based on the assumption that the two samples that are being chosen are independent. However, as mentioned in [3], documents with large number of terms, and greater diversity of terms, are more likely to be retrieved by queries and hence, some document ids in the two samples are likely to be redundant. This may result in violation of the independence condition, which would affect the accuracy of the size estimation.

The Sample-Resample [3] algorithm is the best-known method in the literature, in terms of accuracy and efficiency. It uses terms from the resource description to actually query the database. The resource description (created using the query based sampling technique [4]) is assumed to be present before the estimation is done. The basic assumption behind this technique is that, if the resource description is a good representation of the document collection (i.e. the appearance statistics for each term in the representative corresponds to those in the actual database), the probability of finding a term in the resource description is equal to the probability of finding the term in the database. In other words, let

- D be the number of documents in the actual database*,
- D_T be the number of documents containing term t in the actual database,
- D_R be the total number of documents in the resource description (size of the resource description), and
- D_{RT} be the number of documents containing term t in the resource description,

The Sample-Resample technique assumes that, for any given term t, the condition

$$\frac{D_T}{D} = \frac{D_{RT}}{D_R} \quad (6)$$

holds. Thus, the number of documents in the document collection can then be estimated as follows:

$$D = \frac{(D_T * D_R)}{D_{RT}} \quad (7)$$

The result is averaged over a number of sample queries.

In the ideal case, when the resource description is a very good representative of the actual database, the above assumption is valid because the document frequency of term t in the resource description is proportional to its document frequencies in the actual database. However, in a real life scenario, it is impractical to expect the condition to be satisfied for all terms so that the estimation accuracy would depend on the terms that are chosen to query the database. It is not trivial to find terms that are proportionately represented in the resource description and the actual database, since the search engine is a black box to the meta-search engine. However, experiments show that this method achieves better accuracy than the capture-recapture method. Also, one advantage of this technique is it requires very few queries [3] (as low as five) to probe the database. Thus, its efficiency (the time taken to estimate the size of a search engine which directly depends on the number of probe queries) is much better than the other two techniques mentioned previously.

3. Independence Controlled Sample Size Estimation

In this paper, we propose an approach called *independence controlled sampling,* shown in Figure 1. Our approach makes an assumption different from the one made by the Sample-Resample approach. Our assumption is that the resource description is a good sample of the

* In this paper, the terms "database" and "search engine" are used interchangeably.

1. while (stopping condition not reached)
   1.1. Select a pair of terms from the resource description such that the two terms are independent (The exact method used to check the independence is explained later in Section 3.1).
   1.2. Query the Document Collection for finding the following values: D_1, D_2 and $D_{1\cap 2}$
   1.3. The estimate obtained from this pair is computed as $Est_i = \left(\frac{D_1 * D_2}{D_{1\cap 2}} \right)$ (8)
   1.4. Corrected Estimate (Est_{ic}) = Correction Factor (CF) * Est_i, where,
   CF = (resource description estimate / resource description size)
   $$= \frac{\left(\frac{D_{R1} * D_{R2}}{D_{R1\cap R2}} \right)}{D_R} \quad (9)$$
2. The final estimate is averaged over n individual estimates
   $$Est_f = \frac{\left(\sum Est_{ic} \right)}{n}$$

Figure 1. Search Engine Size Estimation

actual database, as far as the relationship between the terms within the resource description is concerned. That is, terms that are independent in the actual database are also independent in the resource description. Similar to all three approaches mentioned above, we assume that the search engine provides information about the number of documents that match a given query. Based on the above assumptions, our method selects a pair of independent terms from the resource description. The selected terms are then sent to the database, individually and in conjunction, and the number of returned result documents are recorded. The size estimate can then be calculated, by applying probabilistic independence criterion on these numbers.

Some of the terminology used in the algorithm in Fig 1 is mentioned below:

D_1 is the number of documents containing term t_1 in the actual database

D_2 is the number of documents containing term t_2 in the actual database

$D_{1\cap 2}$ is the number of documents containing terms t_1 and t_2 in the actual database

D_R is the number of documents in the resource description

D_{R1} is the number of documents containing term t_1 in the resource description

D_{R2} is the number of documents containing term t_2 in the resource description

$D_{R1\cap R2}$ is the number of documents containing terms t_1 and t_2 in the resource description

There are three points in Figure 1, we would like to elucidate further:

1. In Step 1.1, the algorithm finds two candidate terms that, if independent, can be used to estimate the size of the actual database. We propose two methods to control the independence of the terms namely: (1) the independence criterion in the descriptive statistics and (2) the inferential statistics-based chi-squared test. Once the independent terms are obtained, they are used to estimate the size of the actual database. The two methods will be discussed in section 3.1.
2. The second issue is about an appropriate stopping condition to be used. The stopping condition can either be a simple one like, a predetermined number of term pairs; or, to achieve more accuracy, a slightly complicated condition, which takes into account factors such as the convergence of the average estimate at every iteration. (In this paper, we used the simple condition for our experiments with the number of term pairs fixed at 5).
3. The third issue is a so-called *correction factor* applied to the final estimate in Step 1.4. Since we are using probabilistic statistics to find the independent terms, the two terms thus found may still not be truly independent. This could introduce an error in the final estimate value. A correction factor is used to reduce this error. Since the resource description is assumed to be a good sample of the actual database, the percentage error in estimating the size of resource description and the actual database is assumed to be the same. The two terms found to be independent are used to estimate the size of the resource description. Since the actual size of the resource description is a known piece of information, the correction factor can be computed as the ratio between the estimated size and the actual size of the resource description.

The accuracy of both techniques (Sample-Resample and Independence Controlled Sampling) depends on the faithfulness of the resource description (in a faithful resource description, the probability that a term appears in a document in the resource description should equal the probability that it appears in a document in the actual database). The faithfulness of a resource description cannot be guaranteed in an uncooperative environment. As illustrated in [4], it depends on several factors such as the initial query term, number of query samples, number of documents stored at each stage and so on. Thus, size estimation methods that depend on resource description faithfulness could be critically impacted by

unfaithful resource descriptions. By choosing term independence for estimation and by using a correction factor from the resource description estimate to account for the independence error, our technique is much more flexible and robust to fluctuations in the resource description quality. On the other hand, the accuracy of the Sample-Resample technique is tightly coupled to the faithfulness of the resource description and hence is affected to a greater extent by fluctuations in resource description faithfulness. The above points will be explained in more detail in section 5.1.

3.1. Finding Independent Terms

In this paper, we have used two techniques to check term independence. The first one, which is more primitive, uses the simple descriptive statistics based independence criterion to check if the two terms are independent. For any two terms t_1 and t_2, the independence criterion is specified as follows:

$$\left| P(t_1 \cap t_2) - P(t_1) * P(t_2) \right| < \mu \qquad (10)$$

where, $P(t_1)$ is the probability that a document picked randomly from the sample contains term t_1.
$P(t_2)$ is the probability that a document picked randomly from the sample contains term t_2.
$P(t_1 \cap t_2)$ is the probability that a document picked randomly from the sample contains term t_1 and t_2.
μ is a threshold which is set to a low value.

Table 1. Sample Contingency Table.

		t_2		
		0	1	
t_1	0	f_{00}	f_{01}	*Row_0_total*
	1	f_{10}	f_{11}	*Row_1_total*
		Col_0_total	*Col_1_total*	*Sample_Size*

The second technique is the inferential statistics-based chi-squared test for independence [8]. The relationship between the variables being tested for independence is represented using a contingency table. A sample contingency table is shown above. The chi-squared test of independence is done as follows:

1. State the null and alternative hypothesis. The null hypothesis states that the two terms are independent whereas the alternative hypothesis states that the two terms are not independent.
2. The contingency table (Table 1) is then populated with the observed frequency values (f_{ij}) (number of documents) for each of the cells, from a sample chosen randomly. The subscripts in the observed frequencies denotes the presence or absence of the particular term for example, the value f_{10} would denote the number of documents having term t_1 but not term t_2, the value f_{11} would denote the number of documents with both terms and so on.
3. The next step is to assume independence and compute the expected frequency (e_{ij}) (which must be greater than or equal to five) values from the observed frequency values as follows:

$$e_{ij} = \left[\frac{(Row\_i\_Total) * (Colomun\_j\_Total)}{Sample\_Size} \right] \qquad (11)$$

4. After finding the expected frequency and observed frequency values, the test statistic (χ) is found using the formula:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \qquad (12)$$

The summation is done over all the rows and columns.

5. Once the test statistic is found, based on the required level of significance, it is compared with the test statistic value at the required level of significance (from the chi-squared distribution table). If the obtained statistic is greater, the null hypothesis is rejected, else it is accepted.

4. Experimental Setup

We have conducted experiments to compare the performance of our algorithm with the Sample-Resample algorithm, since the Sample-Resample algorithm performs best among all three existing algorithms that were described in section 2. We also present the results for our technique without the correction factor, in order to illustrate the importance of the correction factor. For the purpose of comparison, we used the same two test beds that were used in [3] which are described below:

1. **Trec-123-100col-bysource**: The test bed contains 100 small databases from the TREC-123 collection. The sizes of the databases are not skewed and the databases themselves are organized by source and publication date.
2. **Trec123-10col:** This test bed was created to test the effectiveness of algorithms on larger databases. For this purpose, ten large databases were created as explained below: The Trec-123-100col-bysource collection was first sorted alphabetically. The first large database was created, by combining every tenth database of Trec-123-100col-bysource starting with the first. The second large database was created, by combining every tenth database starting with the second and so on.

We simulate a search engine (using Boolean Retrieval Model) on every database in each of the test beds.

4.1. Building Resource Description

Resource Descriptions were built for each of the databases in the two test beds using query-by-sampling technique proposed in [4]. Thus, query terms were selected randomly and submitted to a search engine, and the top four documents were retained. This process was carried on until *300* documents were accumulated in the resource description. Resource Descriptions are generally judged based on their *goodness*, which is a complex and abstract notion, difficult to measure. Faithfulness of a resource description is only one of the many factors that contribute towards goodness. In our experiments, we use goodness of a resource description rather than the faithfulness because the goodness of a resource description is a more formal way of evaluating a resource description. Our intuition is that, a good resource description would more or less be a faithful representative of the actual database whereas the vice versa need not be true. Hence, measuring the performance of the estimation algorithms on the basis of goodness instead of faithfulness would yield more comprehensive and reliable results. One measure that has been widely used for measuring resource description goodness is the *Collection Term Frequency ratio* or the *ctf* ratio. It was suggested by [4] for measuring the goodness of a resource description. Essentially, the *ctf* ratio gives a measure of the number of terms in the database that are covered by the resource description. The *ctf* ratio is computed as below:

$$ctf \text{ ratio} = \left(\Sigma_{i \in v^r} ctf_i \right) / \left(\Sigma_{i \in v} ctf_i \right) \quad (13)$$

where ctf_i is the collection term frequency of the term '*i*' (Number of occurrences of the term in the database), *v* is the vocabulary in actual database and v^r is the vocabulary in the resource description. A larger *ctf* ratio denotes better coverage of terms and hence a better resource description.

5. Experimental Results and Analysis

The Mean Average Error Ratio (*MAER*) measure proposed in [3] was used to compare the accuracies of the estimates obtained by the Sample-Resample and Independence Controlled Sampling techniques. The MAER is computed as follows:

$$MAER = mean\left[\frac{(Actual\_Size\_of\_SE) - (Estimated\_size\_of\_SE)}{Actual\_Size\_of\_SE}\right]$$

As can be seen from Table 2, the estimates obtained by the Independence Controlled Sampling approach with correction factor applied is more accurate (at least *10%* better in terms of *MAER*) than the ones obtained by the Sample-Resample for both large databases and small databases. Even without using the correction factor, the Independence Controlled Sampling method obtains better estimates than the Sample-Resample method as can be seen from Table 3. However, the improvements are less significant.

We also studied the effect of the optimality of the resource descriptions on the accuracy of the estimates. For this purpose, the databases were grouped based on the *ctf* ratios of their resource descriptions and the *MAER* for each of these groups was computed. For example, all resource descriptions with *ctf* ratios in the range *0.9* to *1.0* were grouped under one category, those in the range *0.8* to *0.9* fell in another category and so on. The *MAER* for each of the categories for the two methods was then plotted against the *ctf* ratios. Figure 2 shows that the effect of goodness of resource descriptions on Sample-Resample is much more serious whereas its effect on our method (with or without using the correction factor) is less dominant.

Table 2. Accuracy of estimation algorithms based on Mean Average Error Ratio

Collection / Method	Trec-123-100col-bysource	Trec12 3-10col
Sample-Resample	0 .316	0.378
Independence Controlled Sampling (Using independence criterion)	0. 191	0.274
Independence Controlled Sampling (Using chi squared test)	0.192	0.238

Table 3. Accuracy of estimation algorithms based on Mean Average Error Ratio (without CF)

Collection / Method	Trec-123-100col-bysource	Trec12 3-10col
Sample-Resample	0 .316	0.378
Independence Controlled Sampling (Using independence criterion)	0.288	0.294
Independence Controlled Sampling (Using chi squared test)	0.286	0.352

We analyze the above results in detail from two aspects; which are effectiveness (accuracy of estimation),

and, efficiency (number of probe probing queries sent to the database).

5.1. Effectiveness

As far as accuracy is concerned, it can be seen from Table 2 that our method outperforms the Sample-Resample method.

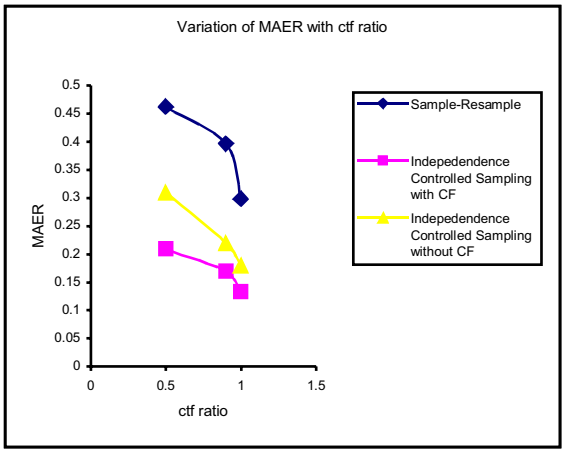

Figure 2: Variation of Mean Average Error with *ctf* ratio.

From Figure 2, we can further observe that the accuracy of the estimate given by the Sample-Resample algorithm depends a great deal on how good the resource description. If *ctf ratio* is taken as the criteria for judging a resource description, then, for smaller databases, resource descriptions generally record most of the terms present in the search engine and hence are fairly accurate. However, as the database size grows, it becomes difficult to build good resource descriptions and the assumption, the document frequency for most terms are same in the resource description and the actual database', becomes weaker and less convincing. As can be seen from Figure 2, for *ctf ratio* values close to 1, both algorithms have very low *MAER,* whereas, at lower *ctf ratio* values, *the* estimates obtained by Sample-Resample begin to deteriorate, while, the deterioration of the estimates for the Independence Controlled Sampling technique is comparatively less severe. Furthermore, as can be seen from Table 2, using the chi-squared technique to test the term independence yields better estimates than the primitive independence criterion test plotted for large databases. A major advantage of our technique is that it is less affected by the quality of a resource description as can be seen from Figure 2 where in, the accuracy of the technique is good even when the ctf *ratio* is low.

The Sample-Resample technique obtains the estimates using statistics from both the resource description and the actual database. Because of this, the technique has no way of finding the error in their estimate in case their assumption is not met. Also, the Sample-Resample technique tends to underestimate the actual database size because, as mentioned in [3], the actual database contains a large vocabulary and the percentage of documents containing a sampled word tends to be overestimated.

On the other hand, the Independence Controlled Sampling method applies term independence to the resource description and finds 'qualified' terms that can then be used to estimate the size of the actual database. The use of term independence facilitates adjusting the final value (Est_i in Figure 1) with a correction factor obtained by finding the error in estimating the size of the resource description (Note that computation of this error requires only the resource description but nothing from actual database). This error is then applied to the final value as a correction. As can be seen from Table 2 and Table 3, the correction factor introduces a significant improvement in the size estimates. This is because, if the size of the resource description is wrongly estimated, then, it is reasonable that the estimates obtained for the actual database will also differ proportionally. Hence, the use of term independence for estimation gives us an intuitive means for correcting certain unbalanced estimates. By combining both term independence control and application of the correction factor, our method provides the all-important robustness, not obtainable by the Sample-Resample method. In other words, compared to the Sample-Resample approach, our approach has better "toughness" or "tenacity" to the environment (in terms of the resource description that is available).

5.2. Efficiency

The improved efficiency of our approach and the Sample-Resample approach, as compared to the other two techniques, is due to the fact that they make effective use of the resource description to choose sample query terms, provided that the resource description is a good representative of the document collection of the search engine. The number of sample queries required by our approach is almost the same as that required by the Sample-Resample approach. Since the cost of querying the search engine is dominant while the local computation costs (i.e. the computation done on resource descriptions) are negligible, it is reasonable to consider the efficiency of our method to be the same as that of the Sample-Resample approach.

6. Conclusion & Future Work

We propose an efficient and effective search engine size estimation technique that outperforms the

existing techniques namely: Interval Estimation, Capture-Recapture and Sample-Resample approaches. This technique takes advantage of the use of resource description to minimize the number of sample queries to be sent to the search engine. It achieves better accuracy by applying a mechanism to select statistically independent term pairs to be used to query search engines and through a mechanism that corrects estimates using data derived from the resource description. All in all, the effect of a sub optimal resource description is less dominant on the Independence Controlled Sampling method as compared to the Sample-Resample method.

In future, we look forward to extending our study to search engines that apply Vector Space Model since only Boolean retrieval model was used in the current experimental setup. We will then further validate our approach by applying it on real time search engines such as university search engines, Google, and AltaVista.

7. Acknowledgement

This work is supported in part by the IT Initiative of the State of Louisiana to Lafayette.

8. Reference

[1] Yuxi Chen, "Statistical Methods to Estimate the Sizes of Search Engines", Technical Report, Computer Science Department, State University of New York at Binghamton, USA,

[2] King-Lup Liu, Clement Yu, Weiyi Meng, Adrian Santoso, C. Zhang, "Discovering the Representative of a Search Engine", Proceedings of Tenth ACM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, 2001, pp.577-579.

[3] Luo Si and Jamie Callan, "Relevant Document Distribution Estimation Method for Resource Selection", Proceedings of 26th annual international ACM SIGIR Conference on Research and development in information retrieval, Toronto, Canada, 2003, pp. 298-305.

[4] Jamie Callan and Margaret Conell, "Query Based Sampling of Text Databases", ACM Transactions on Information Systems (TOIS), New York, USA, 2001, pp. 97-130.

[5] Jamie Callan and Margaret Conell, "Automatic Discovery of Language Models for Text Databases", Proceedings of the ACM SIGMOD Conference, Philadelphia, Pennsylvania, USA, 1999, pp. 479-490.

[6] Jamie Callan, "Distributed Information Retrieval", In W.B. Croft, editor, Advances in Information Retrieval., Kluwer Academic Publishers, Boston, USA, 2000, pp. 127-150.

[7] Michael K. Bergman, "The Deep Web: Surfacing Hidden Value", Journal of Electronic Publishing, University of Michigan Press, Michigan, USA, 2002, pp. 72-78.

[8] Shusaku Tsumoto, "Statistical Independence as Linear Independence", Electronic Notes in Theoretical Computer Science, Department of Medical Informatics, Shimane Medical University, School of Medicine, Shimane, Japan, 2003, Volume 82, Number 4, 12 pages.