Generalising Multiple Capture-Recapture to Non-Uniform Sample Sizes

Paul Thomas
CSIRO ICT Centre
Canberra, Australia
paul.thomas@csiro.au

ABSTRACT

Algorithms in distributed information retrieval often rely on accurate knowledge of the size of a collection. The "multiple capture-recapture" method of Shokouhi et al. is one of the more reliable algorithms for determining collection size, but it relies on samples with a uniform number of documents. Such uniform samples are often hard to obtain in a working system.

A simple generalisation of multiple capture-recapture does not rely on uniform sample sizes. Simulations show it is as accurate as the original method even when sample sizes vary considerably, making it a useful technique in real tools.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*distributed systems*

General Terms

Experimentation, Measurement

Keywords

Size estimation

1. INTRODUCTION

Tools in distributed information retrieval (DIR) use knowledge of a collection's size as a proxy for coverage and completeness, and as important input to algorithms for server selection and language modelling.

Although in some instances servers may report the number of documents they index, many servers do not report a size, and when they do it may be inaccurate or deliberately misleading. DIR tools therefore must generate their own estimates, based on samples of documents acquired with techniques such as query-based [2] or multiple-query [5] sampling.

2. MULTIPLE CAPTURE-RECAPTURE

The "multiple capture-recapture" method (MCR) of Shokouhi et al. [4] builds on Liu et al.'s "capture-recapture" [3]. The central idea for both algorithms is to estimate, given a number of samples from a collection, the expected number of "overlaps"; an "overlap" is when a document is in common between two samples. This depends on N, the number of documents in the collection, so we can use a count of overlaps to estimate N.

If we have T independent samples, each of n documents, then there are $T(T-1)/2$ pairs of samples. Each document has a n/N chance of being in each sample, so a n^2/N^2 chance of being in both; and there are N documents, so we can expect n^2/N overlaps per pair of samples. Overall, the expected number of overlaps $E(O)$ is therefore

$$E(O) = \frac{T(T-1)}{2}\frac{n^2}{N}$$

and if we observe o overlaps, we can estimate

$$\hat{N} = \frac{T(T-1)}{2}\frac{n^2}{o}.$$

It is however often difficult for a working DIR tool to obtain samples of a uniform size n. Of the six sampling methods surved in a recent paper [5], none are able to guarantee samples of a particular size; some, for example the single queries method [1], are extremely unlikely to produce samples of a uniform size. This restricts the use of MCR.

3. A SIMPLE GENERALISATION

A simple generalisation allows MCR to operate with non-uniform sample sizes. For every two samples x and y, we have n_x and n_y, the sizes of each sample, and o_{xy}, the number of documents overlapping in these two samples. T, N, and o are as before. Now

$$\begin{aligned} E(O) &= \sum_{x\in 1\ldots T-1}\sum_{y\in x+1\ldots T} N\frac{n_x n_y}{N^2} \\ &= \left(\sum_{x\in 1\ldots T-1}\sum_{y\in x+1\ldots T} n_x n_y\right)\frac{1}{N};\text{ and} \\ \hat{N} &= \left(\sum_{x\in 1\ldots T-1}\sum_{y\in x+1\ldots T} n_x n_y\right)\frac{1}{o}. \end{aligned}$$

Implementation in a DIR tool is straightforward.

4. VALIDATING THE ALGORITHM

Experiments compared generalised MCR (GMCR) to MCR using simulations with a 1,000,000 "document" collection and random samples. Figure 1 shows the relative errors in 100 runs of each of MCR and GMCR; MCR used samples of 100 "documents" at a time, while GMCR used variable

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

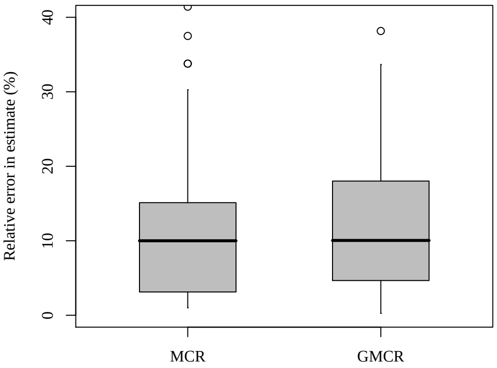

Figure 1: Relative errors in estimation from multiple capture-recapture (MCR), with 100 samples of 100 documents, and generalised MCR (GMCR), with 100 samples of varying size.

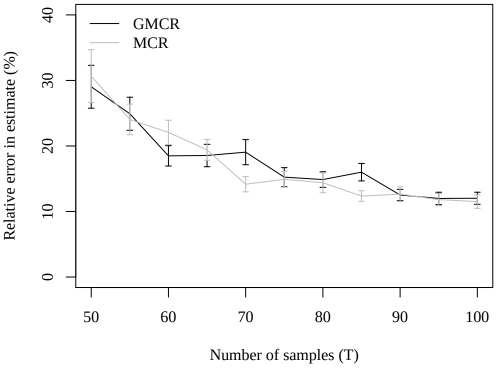

Figure 2: Relative errors in estimation from MCR and GMCR as T, the number of 100-document samples, increases. Bars are ± one standard error.

sample sizes (normally distributed with $\bar{n} = 100$ and $\sigma = 20$). A two-sided t test showed no significant difference in mean relative error; we can conclude that GMCR is as accurate as (although no more accurate than) MCR in this situation. As MCR is one of the best current algorithms, this suggests GMCR is also competitive.

Experiments also examined how the algorithms fared as T, the number of samples, increased. Again, MCR was given 100 documents per sample and GMCR was given a varying number ($\bar{n} = 100$, $\sigma = 20$); T was varied from 50 to 100 with 100 runs at each step. As illustrated in Figure 2, GMCR matched MCR very closely.

Finally, a third set of experiments considered how sensitive GMCR is to variation in sample size. Figure 3 plots relative error, again with 100 samples of a 1,000,000 document collection, as sample sizes become more variable ($\bar{n} = 100$, $\sigma = 1$–40). There is no significant correlation between variability in sample size and relative error, and it appears that GMCR remains equivalent to MCR even if variance in n is high.

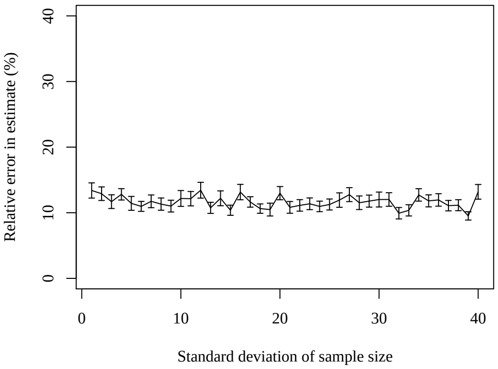

Figure 3: Relative errors in estimation from GMCR as the variability in sample size increases. Bars are ± one standard error.

5. CONCLUSION

The original MCR algorithm is amongst the most accurate size estimation techniques at present, but is of limited use in DIR tools since it requires samples of fixed size. The generalised variant introduced here is as accurate as the original, even if samples vary in size considerably; it therefore seems useful for DIR tools in the general case.

6. REFERENCES

[1] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proc. WWW*, 1998.

[2] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Info. Systems*, 19(2), 2001.

[3] K.-L. Liu, A. Santoso, C. Yu, W. Meng, and C. Zhang. Discovering the representative of a search engine. In *Proc. CIKM*, 2001. Poster.

[4] M. Shokouhi, J. Zobel, F. Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *Proc. ACM SIGIR*, 2006.

[5] P. Thomas and D. Hawking. Evaluating sampling methods for uncooperative collections. In *Proc. ACM SIGIR*, 2007.