# How Much Data Resides in a Web Collection: How to Estimate Size of a Web Collection

Mohammadreza Khelghati, Djoerd Hiemstra, Maurice Van Keulen

## 1. INTRODUCTION

With increasing amount of data in deep web sources (hidden from general search engines behind web forms), accessing this data has gained more attention. In the algorithms applied for this purpose, it is the knowledge of a data source size that enables the algorithms to make accurate decisions in stopping crawling or sampling processes which can be so costly in some cases [4]. The tendency to know the sizes of data sources is increased by the competition among businesses on the Web in which the data coverage is critical. In the context of quality assessment of search engines [2], search engine selection in the federated search engines, and in the resource/collection selection in the distributed search field [6], this information is also helpful. In addition, it can give an insight over some useful statistics for public sectors like governments. In any of these mentioned scenarios, in case of facing a non-cooperative collection which does not publish its information, the size has to be estimated [5]. In this paper, the approaches in literature are categorized and reviewed. The most recent approaches are implemented and compared in a real environment. Finally, four methods based on the modification of the available techniques are introduced and evaluated. In one of the modifications, the estimations from other approaches could be improved ranging from 35 to 65 percent.

*Contributions.* As the first contribution, an experimental comparison among a number of size estimation approaches is performed. Having applied these techniques on a number of real search engines, it is shown which technique can provide more promising results. As the second contribution, a number of modifications to the available approaches are suggested (Table 1 [3]).

## 2. THE SUGGESTED APPROACH

In this work, Heterogeneous and Ranked Model (Mhr), Multiple Capture Recapture (MCR), MCR Regression, Capture History (CH), CH Regression, Generalized Capture Recapture (G-MCR) and Bar-Yossef et al. approaches from the literature are implemented. Having studied these approaches, a number of ideas are suggested to improve their accuracy.

In the approaches like MCR and CH which are based on creating samples and the number of duplicates among them, the idea of considering only the different samples is applied. This can test if different samples can provide more information on the collection size. The similarity of samples is considered as the basic modification idea for MCR and CH.

Different nature of Bar-Yossef et al. needs a different improvement idea. Bar-Yossef et al. is based on a predefined query pool. The number of queries in this pool which cover the collection data is estimated and this number directly affects the collection size estimation. In our experiments over Bar Yossef et al., it was noticed that defining the query pool can highly affect the estimation process. Based on this observation, a different query pool selection method is suggested. In this suggested approach, queries are divided into different query pools based on their frequencies. These pools are indexed and easily accessible by the approach. By sending queries and investigating their results, it is decided if the pool is appropriate or not for the collection. This helps choosing the most appropriate query pool for the collection.

## 3. RESULTS

Having applied the Mhr, MCR, MCR-Regression, CH, CH-Regression and G-MCR approaches on the test set, the results are illustrated in the Figure 1 [3]. These websites are chosen in a way to cover different subjects and have different sizes. In this figure, to be able to compare the performance of the approaches on different data collections of different sizes, the results are normalized by using the Relative Bias metric. If an approach could estimate half of the actual size of a data collection, the corresponding relative bias for that approach is $-0.5$ which is related to $-50$ percent in the figure.

However, it is important to mention that the Bar-Yossef et al. approach implemented in this work was so costly in most of the cases that caused stopping the estimation process. This problem is introduced by the choices of the query pools made during the implementation phase of this approach. Among two pools suggested by Bar-Yossef et al. [1], the one aimed at real cases and not designed for training purposes is implemented. Therefore, the results for Bar-Yossef et al. approach are missing in this part.

*DIR 2013,* April 26, 2013, Delft, The Netherlands.

**Table 1: Improvements Resulting From Modifications**

| | Mhr | MCR | MCR-Reg | CH | CH-Reg | G-MCR |
|---|---|---|---|---|---|---|
| M-Bar-Yossef | 36.25 | 63.67 | 67.36 | 44.74 | 54.70 | 62.77 |
| M-MCR | -19.1 | 8.27 | 11.96 | -10.6 | -0.7 | 7.37 |
| M-MCR-Reg | -24.1 | 3.25 | 6.94 | -15.6 | -5.7 | 2.34 |
| M-CH-1 | 1.35 | 28.77 | 32.46 | 9.84 | 19.79 | 27.86 |
| M-CH-1-Reg | 2.50 | 29.92 | 33.60 | 10.98 | 20.94 | 29.01 |
| M-CH-2 | 0.81 | 28.23 | 31.92 | 9.30 | 19.26 | 27.33 |
| M-CH-2-Reg | 2.77 | 30.19 | 33.87 | 11.25 | 21.21 | 29.28 |

Note: This table provides the percentage of improvements that the modified approaches could result regarding the previously available approaches; considering the average of all the performances on all the tested real data collections on the Web.

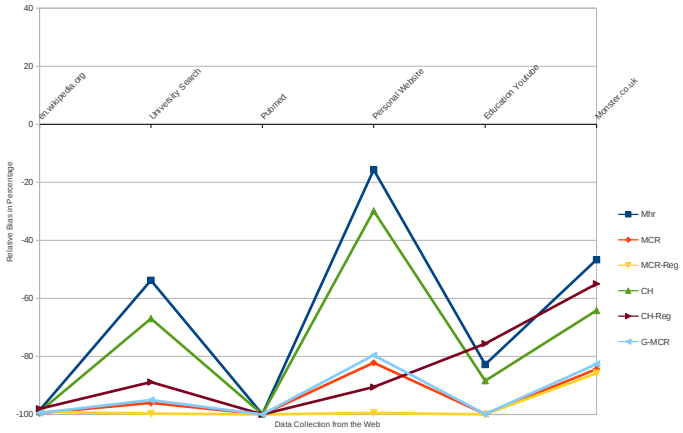


**Figure 1: The Performance of the Approaches on the Real Data Collections from the Web**
Note: The lines are added only to provide more readability of the graph.

## 4. CONCLUSION

Having studied the state-of-the-art in size estimation of non-cooperative websites, the most recent approaches introduced in the literature are implemented in this work. Hence, the MCR, CH, G-MCR, Bar Yossef et al. and regression-based approaches are selected to be studied and compared. To provide an appropriate comparison setting, two issues were regarded highly important. First, the test collection is definrd as a set of websites on the Web from different domains (such as job vacancies, wikis, articles, and personal websites) with different sizes. The second issue was the information available for each approach. The number of sampling events and the samples sizes were set to be the same for all the approaches. Although this test environment could be improved by adding more real deep websites, it is believed that it could provide an appropriate basis for comparing the available size estimation approaches.

Among all the studied approaches, the modified version of Bar-Yossef et al. could provide 35 to 65 percent better estimations on size of the tested deep websites. However, the M-Bar-Yossef et al. approach could not be implemented for the websites which do not provide the access to the content of the search results. In the case of facing such websites, the Mhr approach, both modified versions of the CH approach (M-CH-1 and M-CH-2) and their regressions (M-CH-1-Regression and M-CH-2-Regression) could be among the options to be applied. These approaches had close estimations considering the average performances on all the tested websites.

As future work, we aim at research on the most appropriate time to stop the sampling in estimation process. The alternative approaches could be continuing as far as the limitations or to study questions like what is the adequate number of samples and the most appropriate sample size to provide the most accurate estimation. As another future work, the potential further improvements could be mentioned. As an example, in the selection of pools in the M-Bar-Yossef et al. approach, the selection procedure could be based on the queries from different domains. This classification might lead to higher accuracy of the size estimations.

## 5. ACKNOWLEDGEMENT

This publication was supported by the Dutch national program COMMIT.

## 6. REFERENCES

[1] Bar-Yossef, Z., and Gurevich, M. Efficient search engine measurements. *ACM Trans. Web 5*, 4 (Oct. 2011), 18:1–18:48.

[2] Broder, A. Z., Fontoura, M., Josifovski, V., Kumar, R., Motwani, R., Nabar, S. U., Panigrahy, R., Tomkins, A., and Xu, Y. Estimating corpus size via queries. In *CIKM* (2006), pp. 594–603.

[3] Khelghati, M., Hiemstra, D., and van Keulen, M. Size estimation of non-cooperative data collections. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services* (New York, NY, USA, 2012), IIWAS '12, ACM, pp. 239–246.

[4] Lu, J. Ranking bias in deep web size estimation using capture recapture method. *Data Knowl. Eng. 69*, 8 (Aug. 2010), 866–879.

[5] Shokouhi, M., Zobel, J., Scholer, F., and Tahaghoghi, S. M. M. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR* (2006), pp. 316–323.

[6] Xu, J., Wu, S., and Li, X. Estimating collection size with logistic regression. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 789–790.