Size Estimation of Non-Cooperative Data Collections

Mohammadreza Khelghati
Database Group
University of Twente, Netherlands
s.m.khelghati@utwente.nl

Djoerd Hiemstra
Database Group
University of Twente, Netherlands
d.hiemstra@utwente.nl

Maurice van Keulen
Database Group
University of Twente, Netherlands
m.vankeulen@utwente.nl

With the increasing amount of data in deep web sources (hidden from general search engines behind web forms), accessing this data has gained more attention. This access is limitted to using methods such as crawling and sampling. In the algorithms applied in these methods, it is the knowledge of a data source size that enables the algorithms to make accurate decisions in stopping the crawling or sampling processes which can be so costly in some cases [6]. This 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 [5], search engine selection in the federated search engines, and in the resource/collection selection in the distributed search field [9], 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 the case of facing a non-cooperative collection which does not publish its information, the size has to be estimated **[8]**. Since 1998 that this problem was introduced by Bharat and Broder [4], several techniques are suggested to estimate the sizes of collections on the Web [2, 5, 8, 7, 9]. These estimation approaches root in the techniques applied for human and animals population estimation [1]. In these techniques, the items in collections are sampled through several attempts. Having generated a number of samples, it becomes possible to estimate the size through calculating the similarities and duplicates among the samples [1].

In generating samples for size estimation of document collections on the Web, a query is sent to the search engine, and the returned set of documents is considered as a sample. In selecting the documents to be in this sample, chosen query, content of documents, ranking mechanism and many other features of the collection would effect the probability of a document to be selected. This makes the selection process not random and dependent on a number of factors. In dealing with these factors and mediating their caused biases in the estimation process, different approaches are suggested. In our work, we classify these approaches into three main categories; 1- Approaches which applied bias removal techniques, 2- approaches with both bias removal and random sampling simulation techniques, and 3- No bias removal or sampling simulation apoplied. From these categories, a number of approaches are selected to be evaluated. Multiple Capture Recapture [8], Capture History [8], Bar-Yossef et al. [3] and $Model_{hr}$ [6] approaches are implemented in our experiments. In addition to these approaches, a modified version of Bar-Yossef et al. approach is introduced and implemtented in our work.

In this paper, Having represented a detailed analysis of the introduced approaches, an experimental comparison among them is studied. The suggested approaches in the literature are categorized and reviewed. The most recent approaches are implemented and compared in a real environment. In selecting these real data collections on the Web, it was tried to include different data collections from different domains from which the size of collection could be trusted for evaluation of accuracy of the estimation approaches. Finally, four new methods based on the modification of the available techniques are introduced and evaluated. In one of the introduced methods in this paper, the estimations from other approaches could be improved ranging from 35 to 65 percent. It is also shown what are the shortcomings and faced problems regarding each approach.

Keywords

Deep Web, Size Estimation, Query-Based Sampling, Regression Equations, Stochastic Simulation, Pool-Based Size Estimation, Estimation Bias

1. REFERENCES

[1] Steven Amstrup, Trent McDonald, and Bryan F. Manly. *Handbook of Capture-Recapture Analysis.* Princeton University Press, Princeton, NJ, October 2005.

[2] Ziv Bar-Yossef and Maxim Gurevich. Efficient search engine measurements. *Proceedings of the 16th international conference on World Wide Web*, pages 401–410, 2007.

[3] Ziv Bar-Yossef and Maxim Gurevich. Efficient search engine measurements. *ACM Trans. Web*, 5(4):18:1–18:48, October 2011.

[4] Krishna Bharat and Andrei Broder. A technique for

measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.*, 30:379–388, April 1998.

[5] Andrei Z. Broder, Marcus Fontoura, Vanja Josifovski, Ravi Kumar, Rajeev Motwani, Shubha U. Nabar, Rina Panigrahy, Andrew Tomkins, and Ying Xu. Estimating corpus size via queries. In *CIKM*, pages 594–603, 2006.

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

[7] Jianguo Lu and Dingding Li. Estimating deep web data source size by capture—recapture method. *Inf. Retr.*, 13(1):70–95, February 2010.

[8] Milad Shokouhi, Justin Zobel, Falk Scholer, and Seyed M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR*, pages 316–323, 2006.

[9] Jingfang Xu, Sheng Wu, and Xing Li. Estimating collection size with logistic regression. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 789–790, New York, NY, USA, 2007. ACM.