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

ABSTRACT

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. 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 the crawling or sampling processes which can be so costly in some cases [14]. 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 [7], search engine selection in the federated search engines, and in the resource/collection selection in the distributed search field [19], 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 [17]. In this paper, the suggested approaches for this purpose in the 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.

Keywords

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

1. INTRODUCTION

With the increasing amount of high-quality structured data on the Web, accessing the data in deep web sources have gained more attention. The access to this data is possible through a number of different methods such as crawling and sampling. In the algorithms applied in these methods, it is the knowledge of the data source size that enables the algorithms to make decisions in stopping the crawling or sampling processes which can be so costly in some cases [14]. This tendency is increased by the competition among the businesses on the Web (i.e. jobs and state agencies) to assure their customers of receiving the best possible services [9]. In the context of search engines, the size can highly affect the search engine's quality assessment [7]. Also, in the federated search engines, this information is helpful in the selection of search engines to satisfy the information needs of a posed query. This is also useful in the resource/collection selection in the distributed search [19]. In addition to these advantages, knowing about the size of a data collection can give an insight over some useful statistics which can be interesting for public sectors and governments. For example, having the information from job offering websites can help monitoring job growth in a society [9].

In any of the above mentioned scenarios, in the case of facing a non-cooperative collection which does not publish its information, the size of the collection has to be estimated [17]. In most of the cases, even if the information is published, it could not be trusted. As the only way of accessing these collections is through their query interfaces, the estimator should be able to perform using only a standard query interface. In addition, it should be able to provide accurate estimations, and be applied to any set of documents [7].

Since 1998 that this problem was introduced by Bharat et al. [6], several techniques are recommended through several research work [1, 4, 3, 2, 9, 7, 17, 8, 16, 15, 19]. These techniques can be divided into two main categories; relative size, and absolute size estimators. The relative size estimators provide information on the size of a data collection relatively to the other collections while the absolute size estimators estimate the absolute size of a collection. From the first category, the work in [6] and [12] could be considered. The approaches introduced in the second category could be further classified based on a number of different technical aspects. These classifications are described in the following.

Need Documents Content or Documents IDs. First, the approaches which need analyzing the content of the selected

iiWAS2012, 3-5 December, 2012, Bali, Indonesia.

documents, and second, the approaches which only need to know about the IDs of the selected documents.

How to Deal with Bias. The approaches introduced for the collection size estimation are based on the Query-Based Sampling (QBS). In the QBS, by sending a query to the search engine, a set of documents would be sampled [16, 6]. In this approach, it is assumed that the document samples are generated randomly, while in reality, the chosen query, the content of documents, the ranking mechanism and many other factors would affect the probability of a document to be selected. This makes the selection process not random and could introduce biases in the estimations. Based on the methods applied to resolve this situation, three different categories are suggested to be applied;

1. The approaches which use the techniques to simulate the random sampling and get closer to a set of randomly generated samples (i.e. Bar-Yossef et al. approach **[4, 3]**). They also apply techniques to prevent and remove bias.

2. There are also a number of approaches which accept the non-randomness of the generated samples and try to remove the known biases (Mhr [15], Multiple Capture-Recapture Regression [17], Capture History Regression [17], and Heterogeneous Capture [19]).

3. In this category, there are approaches that accept the samples as they are and do not try to remove the possible biases. These approaches are highly potential to produce biases in the estimations (Sample Resample, Capture History, Multiple Capture Recapture [17] and Generalized Multiple Capture Recapture [18]).

From each one of these categories, there are a number of different approaches applicable for estimating the size of a non-cooperative website. This arises the challenge of the most appropriate approach selection to apply for a deep website available on the Web. To our best knowledge, there is no thorough performance comparison among these available techniques. The availability of such a comparison could also determine if there is still space for further improvements. A general overview on the issues mentioned in this section is illustrated in Figure 1.

Contributions. As the first contribution of this work, an experimental comparison among a number of size estimation approaches is performed. Having applied these size estimation techniques on a number of real search engines, it is shown that which technique can provide more promising results and what are the shortcomings and faced problems. As the second contribution, in addition to this experimental study, a number of modifications to the available approaches are suggested in this work. The amount of improvements which these modifications could bring along to the previous versions are also provided in this paper.

Structure of the Paper. In the next section, a number of approaches from each one of the three categories (categorized based on the way dealing with bias) are introduced and discussed. This study provides a solid basis for performing the experiments which is described in Section 3 (the Experiments Section). In this section, the selected available approaches The possible improvements to each one of the tested estimation approaches are discussed and illustrated in Section 4 (Improvements Section). The results of these experiments and the analysis over these results are represented also in Section 4. Finally, the conclusion and future work are mentioned in Section 5.

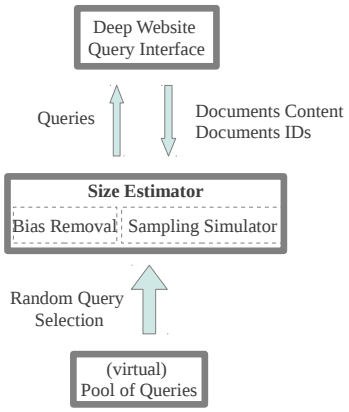

Figure 1: The General Overview on Data Collection Size Estimators.

2. BACKGROUND

The data collection size estimation approaches root in the techniques applied for human and animals population estimation[1]. These methods are based on the ratio between the known (marked) and unknown (unmarked) parts of a collection. In the domain of the deep websites, these approaches could be classified into two categories; the absolute and the relative size estimators. The approaches in the second category could be further divided into three classes based on the methods applied for dealing with the introduced biases. In the following of this section, sample approaches from each one of these three classes are mentioned.

2.1 Approaches Accepting Samples As-They-Are and no Bias Removal

Sample Resample Approach. In the Sample Resample approach (SRS), having the initial query from a list of terms [8, 17], the next queries are selected at random from one of the retrieved documents by the previous query. This sampling process stops after downloading a predefined number of documents. The document frequencies of a term in the sampled documents and in the collection provide an estimation of the size of that collection [17].

Capture Recapture Approach. The Capture Recapture method has roots in ecology and is based on the number of duplicates among different captured samples. In order to estimating the size of a type of animals, for example tigers, at first, a number of tigers would be captured, marked and

[1]France Human population estimation by Pierre Laplace (1749–1827) and earlier applications for fish and duck populations [1].

released. After a while, in another try, another number of tigers would be captured. By counting the duplicates in these two samples and using 1 Equation, it would be possible to estimate the size of the collection [1]. In this formula, if the two samples are not big enough to have any duplicates, it could not be possible to have any results. As a solution, multiple and weighted capture-recapture methods are introduced which would be explained in the following subsections. The application of this technique in the data collection size estimation was first introduced by Liu et al. [16]. In the Liu et al. work [16], it is not described how to implement the proposed approach in practice. It is unclear what the sample size should be and how a random sample might be chosen from a non-cooperative collection. In the traditional Capture Recapture, if two samples are not big enough to have any duplicates, it is impossible to have any results. As a solution, multiple and weighted capture-recapture methods are introduced.

$$\frac{firstSampleSize \times secondSampleSize}{duplicates} \quad (1)$$

Multiple Capture Recapture (MCR). To resolve the issues in traditional capture recapture, a weighted method was introduced in the Shokouhi et al. work [17]. In this approach, by gathering the T random samples of size m, and counting duplicates within each sample pair, the expected size of the collection is computed. Only the document identifiers are required for estimating the size of the collections.

Generalized Multiple Capture Recapture (GMCR). In the application of the MCR, it is necessary to have the samples of the same size. However, it is often difficult to obtain samples of a uniform size. This restricts the use of the MCR. In the work by Thomas [18], a generalization over MCR is suggested to enable it operating also with samples of non-uniform sizes.

Capture History (CH). In the Shokouhi et al. work [17], a weighting function is introduced for the capture recapture technique. This approach is called the Capture History (CH) and utilizes the total number of documents in a sample, the number of documents in that sample that were already marked, and the number of marked documents gathered prior to the most recent sample to introduce a weight for the sample. In the CH approach, it is assumed that the probability distribution of each individual satisfies a uniform distribution. However, this is not the case in search engines and causes biases in the estimation results.

Broder et al. Approach - Extra Pool. In the "Estimating Corpus Size via Queries" [7], two approaches are introduced which are both based on a basic estimator. In the basic estimator, the weight of a document is defined as the inverse of the number of terms in the document which are also in the pool[2]. Accordingly, the weight of a query is defined as the sum of the weights of all the documents containing that query. By calculating the average of several query weights, an approximation to the basic estimator is obtained. The first approach belongs to the third category and is described later. In the second approach, two query pools covering two independent subsets of the corpus are needed. In this context, independence does not mean disjointness. It means that they may share documents, but fraction of documents that belong to one pool should be the same whether we consider the entire corpus or just the other pool [7]. This approach estimates only the part of the corpus in which the pools are uncorrelated. In practice, it might be hard to obtain such sets of queries.

[2]A precomputed pool of uniformly sampleable queries

2.2 Approaches Based On Removing Bias

In applying the QBS, different factors like the chosen query, document properties, and the search engine's specifications could affect the sampling process. Detecting all these factors and resolving them could be so costly or even not possible in some cases [3, 15]. Therefore, some approaches focus on removing the biases caused by these factors. In the Bharat et al. work [6], two major biases are introduced; query bias and ranking bias. The query bias addresses the fact that for different queries, documents would have different chances to be chosen. Returning only the top-k results, and ranking algorithms applied in search engines could also cause bias in the size estimation. This is known as the ranking bias.

Regression Equations. Regression analysis is a statistical tool used for estimating a variable which is dependent upon a number of independent variables [13]. It investigates the relationships between these variables and also provides the degree of confidence that the prediction is close to the actual value. In regression analysis, the variation in the dependent variable is represented by a value shown as R^2. The R^2 value is between zero and one, and shows to what extent the total variation of the variable is explained by the regression. A high value of R^2 suggests that the regression model explains the variation in the dependent variable well. In regression analysis, omitted variables and closely-correlated independent variables (if their effects are difficult to separate) could make problems in the estimation process [13]. As mentioned in the Capture History (CH) and Multiple Capture Recapture (MCR) Subsections, the MCR and CH approaches introduce biases in the estimations which lead to underestimating the collections sizes. To compensate for the selection bias in these approaches, the relationship between the estimated and absolute collection size is approximated by the regression equations. These approaches are referred as the MCR-Regression and CH-Regression. In another approach, by Xu et al., called the Heterogeneous Capture [19], the capture probabilities of documents in the sampling process is modeled with logistic regression. In calculating these probabilities, the document and query characteristics are modeled as a linear logistic model.

Heterogeneous And Ranked Model (Mhr). In the work by Lu [14], he introduces a model to reduce the ranking bias based on a previous work in Lu et al. work [15] for removing the query bias. In the Lu et al. work [15], with the assumption of having random samples from a uniform distribution, an equation between the overlapping rate and the percentage of examined data is suggested. In this equation, the overlapping rate is defined as the total number of all documents divided by the number of distinct documents cached during the sampling procedure. By relating this overlapping

rate to the capture probability of a document in any of the iterations of sampling, and applying linear regression, the Heterogeneous Model (Mh) is introduced. Lu et al. mention that this method can resolve the query bias and can be only applicable to the search engines that do not produce overflowing queries [15]. The overflowing queries are the queries for which the matched results are more than the returned results. This problem is addressed in Lu's work [14] by multiplying the model introduced in Lu et al. work [15] by overflowing rate of queries. Overflowing rate represents the total number of matched documents for a query by the total number of returned documents for that query. This model is named as the Heterogeneous and Ranked Model (Mhr). If the total number of answers returned for a query (matched documents) and the number of results that user can view from all the matched documents is not available, the model would be reduced to the Mh model [15].

2.3 Having Close-To-Random Samples And Bias Removal

To have the random or close-to-random samples, one of the possible techniques is the stochastic simulation methods like Monte Carlo algorithms [11]. From the Monte Carlo simulation methods, rejection sampling, importance sampling and metropolis-hastings methods are applied in the research work in this field [3, 4]. The first two would be explained in the following paragraphs.

These techniques are based on producing biased samples with corresponding weights of sampled documents representing their capture probabilities. The availability of these weights allow the application of the stochastic simulation methods [3]. The stochastic simulation techniques accept the samples from a trial distribution ($Q(x)$) and simulate sampling from a target distribution ($P(x)$). Therefore, by defining a $Q(x)$ which has uniform distribution and is easily sampled, the unbiased sampling could be done for the $P(x)$ [11]. In the rejection sampling, it is assumed there is a $Q(x)$ with a predefined constant c that $P(x) < c * Q(x)$. Having generated samples from the $Q(x)$, the samples would be in the $P(x)$ if they satisfy that inequality [11]. In the rejection sampling, the samples do not belong to the $P(x)$. In the importance sampling, instead of generating samples from a probability distribution, it is focused on estimating the expectation of a function under that distribution [11]. For each generated sample, a weight is also introduced. This weight is used to represent the importance of each sample in the estimator.

Broder et al. Approach - Sampling. As mentioned before, Broder et al. introduce two approaches [7]. In the sampling approach, the size is estimated through using the size of the pool, the basic estimator, and the ratio between the number of documents represented by the queries in the pool and the collection size. As this can be so costly, the ratio is estimated by sampling documents. Calculating the weights for queries in this method implies that the approach is implicitly using the importance sampling [4]. In this approach, it is not analyzed how the presence of the degree mismatch (the difference between the predicted document weight and the actual one) can cause bias [4]. Also, removing the overflowing queries from the pool may incur missing the queries which should be considered and including additional queries which should not be considered [4].

Table 1: Test Set - Real Data Collections on the Web

Data Collection	Size* (number of documents)
A Personal Website *http://wwwhome.cs.utwente.nl/~hiemstra/*	382
University Search Website *http://www.searchuniversity.com/*	4,076
Job Search Website *http://www.monster.co.uk/*	40,000**
Youtube Education *http://www.youtube.com/education/*	311,000
English corpus of Wikipedia *http://en.wikipedia.org/*	3,930,041
US National Library of Medicine - English Documents *http://www.ncbi.nlm.nih.gov/pubmed/*	17,606,509

* The collections sizes are reported on 12/7/2012.
** Although the size is not published, this is a close estimate by browsing jobs in sections.

Bar-Yossef et al. Approach. To resolve the issues in the Broder et al. approach, a new method is suggested by Bar Yossef et al. [4]. In this work, the sample space is defined as a pair of a query and a document (q, d). This eliminates the need to use the rejection sampling for the random selection of the queries [3]. Instead of sampling from the target distribution, the estimator samples a document from a different trial distribution which allows easier random sampling. Having submitted a number of randomly selected queries to the search engine, valid results for each query would be determined. A valid result for a query is defined as a document which is returned by the search engine for the query and also contains that query. The procedure stops when reaching a query that has at least one valid result. Considering this as a sample, with the number of documents in the valid graph for the sample query, the estimation of the inverse document degree, and the size of the valid queries pool (estimated through the random sampling), the size of the collection could be estimated. The inverse degree estimation is performed by submitting the terms in the content of the page which are also in the pool to the search engine. If that page is among the submitted query's results, the procedure stops [4].

3. EXPERIMENTS

As one of the contributions of this paper, an empirical study is performed on the suggested approaches for estimating the sizes of the collections. In this study, these approaches are applied to real cases; data collections available on the Web. In selecting these data collections, it was tried to include collections with different sizes and from different subject areas. In the Table 1, a list of the data collections used in this experiment and their corresponding sizes are illustrated.

Although implementing all the approaches could be an ideal situation, due to the lack of time and resources, only a num-

ber of introduced approaches are implemented in this work. In the selection of the approaches, we attempted to cover all three categories in the absolute size estimators. Therefore, the MCR, MCR-Regression, CH, CH-Regression, Mhr and the Bar-Yossef et al. approaches are chosen to be implemented.

Implementation Differences. It is important to point out that in the implementations of the approaches in this paper, if there are no duplicates found among the samples, the number of duplicates is set to be 1. This enables the approaches to provide an estimation even without any duplicates. In addition, for the MCR approach, samples of a fixed size are needed. Therefore, the average size of all the samples is set as the sample size in the calculation.

Performance Measure. In order to get a more accurate performance of an approach, each approach is repeated 100 times for a predefined sample size, and the number of samplings. The results of these iterations are represented by the Relative Bias (RB) [15]. The RB measures how close the estimations are to the actual size and is calculated through Equation 2. In this formula, $E(\overline{N}) = \frac{\overline{N_1}+\overline{N_2}+...+\overline{N_{Times}}}{Times}$ represents the mean value of the $Times$ number of estimations.

$$RB = \frac{E(\overline{N}) - N}{N} \quad (2)$$

In some parts of the charts using the RB measure, the comparison of the approaches become not so clear. Therefore, to provide more clear performance comparison charts, another measure is also used. As it is shown in Figure 3, the $Log10(E(\overline{N})/N)$ is calculated for the approaches. These two measures could provide a better overview on the performance of the approaches and make it easier to do the comparisons.

3.1 Applied Query Pools

To be able to apply the introduced techniques in this paper, three different pools of queries are developed; pool A, pool B and pool C. In creating the pools, it is tried to follow the requirements mentioned in the papers for each one of the approaches. The query pool A is developed for the Mhr, CH, MCR, G-MCR and the regressions approaches. It includes the top 1000 most frequent words extracted from the pages of the Wikipedia and the pages included in the ClueWeb09 Dataset[3].

The next query pool, the pool B, is designed to be similar to one of the pools mentioned in the paper by Bar Yossef et al. [5]. In that paper, two different pools are used for sampling. The first one is for the training purposes and is a pool of 43 million phrase queries of length 4, extracted from the pages in part of ODP data set [10]. However, in order to run the approach for estimating the sizes of the real cases on the Web, a different pool is applied. Based on the specification of this pool, the pool B is created with 2.775 billion queries. It includes 1.5 billion decimal strings of 5 to 9 digits, 7.4 million single terms extracted from the Wikipedia website, 18 million single terms extracted from the ClueWeb09 Dataset, and 1.25 billion two-term conjunctions of the 50,000 most

[3]http://lemurproject.org/clueweb09.php/

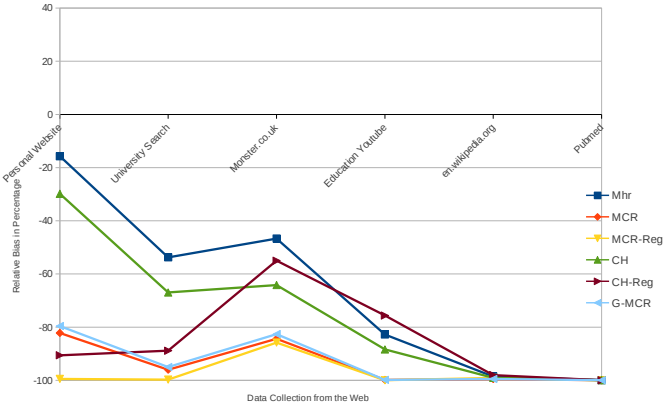

Figure 2: The Performance of the Approaches on the Real Data Collections on the Web. The lines are added only to provide more readability of the graph. The closer the points are to y=0, the more accurate estimations they represent.

frequent extracted single terms (excluding the 100 most frequent ones).

The third pool, pool C, is applied for the M-Bar-Yossef approach which is introduced in the 4.1 Subsection. This pool consists of four different pools; pool-a, pool-b, pool-c, and pool-d. These pools sequentially consist of the most 10^4, 10^5, 10^6, and 10^7 frequent terms extracted from the pages retrieved from the Wikipedia and parts of the Web (ClueWeb09 Dataset). In addition, for each pool, for the same amount of terms, integer digits are added. Therefore, the pools are of the sizes of 2×10^4, 2×10^5, 2×10^6, and 2×10^7.

3.2 Results

Having applied the Mhr, MCR, MCR-Regression, CH, CH-Regression and G-MCR approaches on the test selection of real websites, the results are illustrated in the Figure 2. In this figure, to be able to compare the performance of the approaches on different data collections with different sizes, the results are normalized by using the Relative Bias metric. If an approach could estimate the 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.

There are two points which should be clarified in this part. First, the Bar-Yossef et al. approach implemented in this work was so costly in the 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. [5], 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. There is a solution for this problem which would be introduced in the Improvements Section of this report. Second, as mentioned in the Background Section, the Mhr approach would be changed to the Heterogeneous Model (Mh) in the absence of information on the numbers of matched and returned documents. However, in this report,

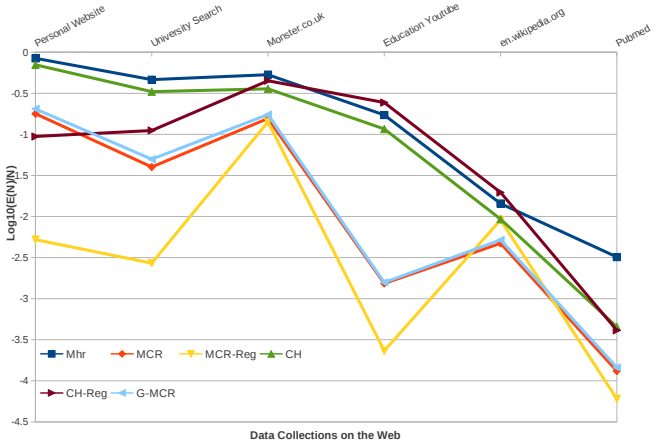

Figure 3: The Performance of the Approaches on the Real Data Collections on the Web. The lines are added only to provide more readability of the graph. The closer the points are to y=0, the more accurate estimations they represent.

as the algorithm of the Mhr approach is implemented, the implementation is still referred as the Mhr approach.

4. IMPROVEMENTS

As the second contribution of this paper, we attempted to improve the performance of the introduced size estimation approaches in the Background Section. As a result, four different approaches are suggested based on applying modifications on the available versions. These approaches are described in details in the following subsections.

4.1 Modified Bar-Yossef (M-Bar-Yossef)

The performance of the Bar-Yossef et al. approach depends highly on the selection of the query pool. In our experiments, the setting of a big pool resulted in so costly processes which could not be followed due to the limitations. It was also observed that if the pool selection is done in a better way, the results would outpass the other introduced approaches. In this context, a better pool could be defined as a pool which covers more pages in the document collection. To provide such a pool for each document collection with different features, it seems necessary to adopt different pools. This difference could be the number or the type of the queries included in the pool. An intelligent pool selection is the main idea to improve the results of the Bar-Yossef et al. approach and make it possible to be applied for document collections of different sizes. This means that the size and coverage of the pool are increased step by step based on the feedback obtained from previous queries posed on the document collection.

In our suggested approach to improve the performance of the Bar-Yossef et al. method, it is suggested to start with a small pool. After sending a small number of queries, based on the number of found queries and documents in the valid graph (described in 2.3 Subsection), it is decided if a bigger pool would serve this estimation better. Then, the bigger pool is selected and processed. The features of these pools are described in detail in the Experiments Section. This pool selection process continues till the best pool or the biggest pool is reached. It is important to point out that the previously sent queries and found results are used in the next phase and the pools are already indexed. These issues make it possible to implement the intelligent pool selection without any extra cost.

The performance of the Modified Bar-Yossef approach which is referred as M-Bar-Yossef is illustrated in the Figure 4 (RB) and Figure 5 ($Log10(E(\overline{N})/N)$). The improvements introduced by this approach to the results of other tested approaches over the size estimation of all the tested real data collections (listed in the Table 1) are illustrated in the Table 2. In this table, the average performance of approaches on all the data collections is considered. As it is shown in this table, the M-Bar-Yossef approach could provide 35 to 65 percent closer estimations considering all the tested deep websites.

4.2 Modified Multiple Capture Recapture (M-MCR)

The modification applied to the MCR method is based on the idea that different samples could be a better source of information for the size estimation process. To test this idea, for improving the performance of the MCR method, similar samples are removed. The similarity between two samples is judged based on the number of duplicates between those samples. This similarity threshold should be adjusted beforehand. In this work, it is set as the 30 percent of the sample size. This modification is referred as Modified MCR (M-MCR). The average performance of the M-MCR approach in comparison to the other introduced approaches is shown in the Table 2. In the Figure 4 (RB) and Figure 5 ($Log10(E(\overline{N})/N)$), it is possible to compare its performance with all the other introduced approaches. This modification could be also helpful in decision on when to stop the sampling process. Although this issue is not studied in this work, it could be considered as a future work. The estimated size by this modified version of the MCR is used by the regression formula introduced in the MCR-Regression. This is referred as the M-MCR-Regression approach. The results and improvements of this approach are shown in the Figure 4 (RB), Figure 5 ($Log10(E(\overline{N})/N)$) and the Table 2.

4.3 Modified Capture History (M-CH-1)

The approach introduced in this work to improve the CH approach is based on the same idea mentioned in the M-MCR; removing the similar samples by counting the duplicates between pairs of samples. If the number of duplicates between two samples is more than 30 percent of the sample size, two samples are judged to be similar and only the earlier found sample would be included in the calculations. As mentioned in the previous subsection, this could provide information on the time to stop the sampling process. As there are two modifications introduced for the CH approach, this modification is called as Modified CH-1 (M-CH-1). The results of the average performance of the M-CH-1 in comparison to the other introduced approaches are provided in the Table 2. In the Figure 4, it is possible to compare the performance of the M-CH-1 with all the other approaches. The improvements this approach could introduce to the size estimation of the document collections in comparison to the

Table 2: Improvements Resulting From the 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.

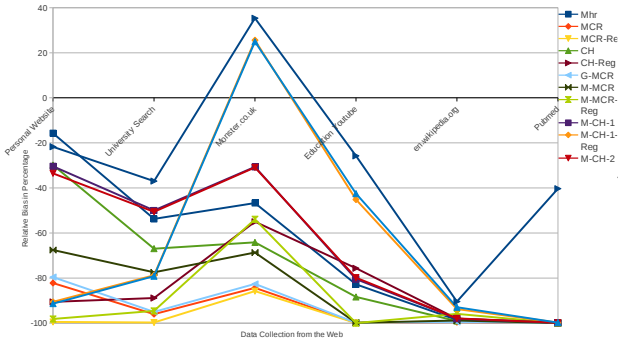

Figure 4: The Performance of All the Approaches on the Real Data Collections on the Web. The lines are added only to provide more readability of the graph. The closer the points are to y=0, the more accurate estimations they represent.

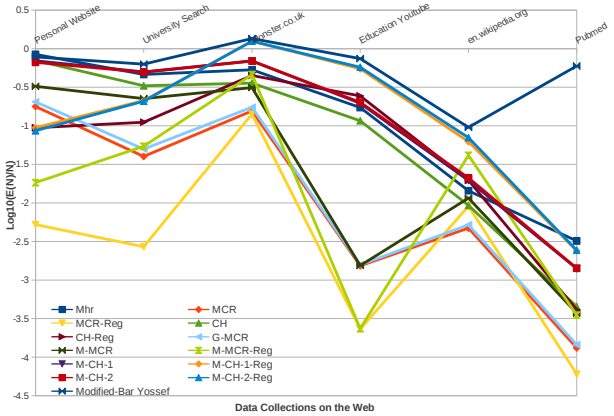

Figure 5: The Performance of All the Approaches on the Real Data Collections on the Web. The lines are added only to provide more readability of the graph. The closer the points are to y=0, the more accurate estimations they represent.

other tested approaches are mentioned in the Table 2. The estimated size by this modified version of the CH is used by the regression formula introduced in CH-Regression. This is referred as the M-CH-1-Regression approach. The results and improvements of this approach are shown in the Figure 4 (RB), Figure 5 ($Log10(E(\overline{N})/N)$) and the Table 2.

4.4 Modified Capture History (M-CH-2)

As another approach to improve the performance of the CH, the similar samples are judged based on the number of duplicates in the sample considering all the previously captured documents. If this number is more than 50 percent of the sample size, the sample is not included in the calculations. This information could be also helpful in the decision on stopping the sampling process. This modification is referred as Modified CH-2 (M-CH-2). The results of its average performance over all the real data collections are shown in the Table 2. In the Figure 4, it is possible to compare the performance of the M-CH-2 with all the other approaches introduced in this paper. The estimated size by the M-CH-2 is used by the regression formula introduced in the CH-Regression. This is referred as the M-CH-2-Regression approach. The results and improvements of this approach are shown in the Figure 4 (RB), Figure 5 ($Log10(E(\overline{N})/N)$) and the Table 2.

5. CONCLUSION AND FUTURE WORK

Having studied the state-of-the-art in the size estimation of the non-cooperative websites, the most recent approaches introduced in the literature are implemented in this work. Hence, the Multiple Capture Recapture, Capture History, Generalized Multiple Capture Recapture, Bar Yossef et al. and regression-based approaches are selected to be studied and compared. To provide an appropriate comparison environment, two issues were regarded highly important. First, it was decided to provide 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. This made it possible to observe the performance of each approach with the same available data. 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 the Bar-Yossef et al. approach could provide 35 to 65 percent better estimations on the size of the tested deep websites on the Web. 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.

While a wide detailed overview could be provided on the available techniques and approaches applied in the size estimation of the non-cooperative websites in this research work, a number of questions are still without solutions. As one of the most important issues in the size estimation process, it is not yet known what is the most appropriate time to stop the sampling process. As one of the strategies, having in mind that all the estimation approaches provided better results with more available data, continuing as far as the limitations permit is one of the options. The other alternative is 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. The existing approaches could be further improved by conducting more studies in this area. 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.

6. ACKNOWLEDGEMENT

This publication was supported by the Dutch national program COMMIT.

7. REFERENCES

[1] Amstrup, S., McDonald, T., and Manly, B. F. *Handbook of Capture-Recapture Analysis*. Princeton University Press, Princeton, NJ, Oct. 2005.

[2] Anagnostopoulos, A., Broder, A. Z., and Carmel, D. Sampling search-engine results. In *WWW '05: Proceedings of the 14th international conference on World Wide Web* (New York, NY, USA, 2005), ACM Press, pp. 245–256.

[3] Bar-Yossef, Z., and Gurevich, M. Random sampling from a search engine's index. In *Proceedings of the 15th international conference on World Wide Web* (New York, NY, USA, 2006), WWW '06, ACM, pp. 367–376.

[4] Bar-Yossef, Z., and Gurevich, M. Efficient search engine measurements. *Proceedings of the 16th international conference on World Wide Web* (2007), 401–410.

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

[6] Bharat, K., and Broder, A. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst. 30* (April 1998), 379–388.

[7] 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.

[8] Callan, J. P., and Connell, M. E. Query-based sampling of text databases. *ACM Trans. Inf. Syst. 19*, 2 (2001), 97–130.

[9] Dasgupta, A., Jin, X., Jewell, B., Zhang, N., and Das, G. Unbiased estimation of size and other aggregates over hidden web databases. In *Proceedings of the 2010 international conference on Management of data* (New York, NY, USA, 2010), SIGMOD '10, ACM, pp. 855–866.

[10] dmoz. `http://dmoz.org`, Title=The open directory project.

[11] Geiger, D., Heckerman, D., and Meek, C. Introduction to monte carlo methods. In *Learning in graphical models*, M. Jordan, Ed. Kluwer, 1998, pp. 175–289.

[12] Gulli, A., and Signorini, A. The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (New York, NY, USA, 2005), WWW '05, ACM, pp. 902–903.

[13] Kern, J. C. An introduction to regression analysis. *American Statistician 61*, 1 (2007), 101–101.

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

[15] Lu, J., and Li, D. Estimating deep web data source size by capture—recapture method. *Inf. Retr. 13*, 1 (Feb. 2010), 70–95.

[16] lup Liu, K., Yu, C., and Meng, W. Discovering the representative of a search engine. In *In Proc. CIKM* (2001).

[17] 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.

[18] Thomas, P. Generalising multiple capture-recapture to non-uniform sample sizes. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 839–840.

[19] 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.