

Estimation Methods for the Size of Deep Web Textural Data Source: A Survey

Jie Liang
School of Computer Science
University of Windsor

The estimation of the size of deep web data sources has been an open problem since 1998. This survey reviews all papers that were available online, and other, resources, on estimating the size of data sources during the period 1998 to 2008. In the survey, we first clarify several basic terms that are used in the survey but whose meanings vary in the literature. Basic models in the literature on estimation are also discussed. The survey introduces query-based sampling approaches and reviews the estimation methods of estimating relative size and actual size of data source(s). Query-based sampling is biased. The survey also reviews research on overcoming biases caused by various estimation methods. Finally, the future direction of estimation is discussed.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

General Terms: estimation methods, size, capture-recapture, query-based sampling, overlapping
Additional Key Words and Phrases: multi capture-recapture, regression analysis, document weight, query weight, overlapping rate

Contents

1	Introduction	2
2	Background	2
3	The Basic Estimation Method	4
3.1	Query-based sampling	4
3.2	Identifying documents	5
3.3	Estimating relative size	5
3.4	Estimating actual size	6
3.4.1	Methods based on document IDs	7
3.4.2	Methods based on document content analysis	9
4	Biases	13
4.1	Query Bias	14
4.2	Rank Bias	14
5	Overcoming Biases	14
5.1	Methods to Overcome Query Bias	14
5.2	Methods to Overcome Ranking Bias	15
6	Other Estimation Methods	15
7	Conclusion and Future Work	16

1. INTRODUCTION

Research on estimating the size of deep web data sources has been conducted for almost 20 years. In the literature, the development of research in this area can be separated into two stages: estimating the relative size of deep web data sources and actual size estimation.

From the 1980's, researchers started to estimate several metrics of the fast growing web, like the total number of sites in the web, it's physical size and its content classification. At the very beginning, research started to estimate the size of the web via search engines. Five papers introduced in the survey describe work on the relative search engines size estimation. And based on these papers, the size of the web could be inferred. A number of probability models used in many other estimation methods were proposed in this stage.

Although estimating the size of web has started from the first stage, the results were far from accurate. After a debate on the information released by two large search engine companies, Yahoo! and Google, research attention changed to estimate the actual size of deep web data sources. Enlightened by the research in biostatistics, researchers in Computer Science proposed Capture-Recapture methods and many improved methods that are base on this methods. Sampling deep web data sources causes bias. In the second stage (2003-2008), much attention was directed at reducing the bias caused by query-based sampling technique. This survey studied eleven papers worked on estimating the actual size, including books, reports and papers.

The survey is organized as follows: Section 2 briefly reviews the background of research on estimation of the size of data collections. It also clarifies the basic terminologies in this field and introduces the basic probability model that almost all methods are based on. Section 3 studies the basic sampling method that was used in many methods. In section 3, we review all methods proposed in the literature on estimating relative size and actual size. Estimating the actual size is discussed in two branches in the literature. Section 4 discusses the bias which exists in the methods proposed. Section 5 briefly reviews the methods to overcome those biases discussed in Section 4. Section 5 introduces two non-traditional methods to estimate the size of deep web data source. Section 6 summarizes the survey and presents a possible improvement to existing estimation methods.

2. BACKGROUND

The deep web, in contrast to surface web, usually includes those resources that can only be obtained via query interfaces ("across" surface web) [Bergman 2001]. Sometimes it refers to searchable databases. In the literature, there are many terms referring to the textual data source, like "data collection", "document archive", and simply, "database" and more general, "corpus". Hence, the term *deep web data source* refers to those data sources that can only be accessed via query interfaces. In most of the papers included in the survey, the researchers consider the search engine as a large deep web data source while the whole web is the largest one. However, research is not limited to search engines. Other deep web data sources, identified

by the features of the deep web, include the background databases of many E-commerce web sites, for example, Amazon.com [Lu 2007] and dangdang.com [Ling et al. 2008].

In this survey, the term *size* of deep web textual data source refers to *the total number of documents* or *the total number of pages*. Many papers in the literature also propose methods to estimate the physical size (in gigabytes, etc.) of a data source. Although estimating physical size has also been widely studied, this survey does not draw much attention to it.

Estimating the size of data source is a necessary step of a data crawler and data extractor. They need to know the size to decide when to stop crawling and to evaluate the performance of the crawler and the extractor [Bharat and Broder 1998][KISHI et al. 2000][Broder et al. 2006][Lu 2007]. Also, The size of the data source is an important evaluation metric of search engines [Bergman 2001][Callan and Connell 2001][Bar-Yossef and Gurevich 2006]. At the early stage of research on the web (1990-2000), research focused on estimating how large the web is. Many researchers developed approaches to estimate the *relative size* in order to evaluate which search engines has a larger coverage of the web. Also, the size of the web can be inferred. In 2005, two large search engine companies released their coverage of the web [Price 2005]. However, people questioned their information and researchers started to work on estimating the *actual size* of deep web data source.

Some estimation methods originate from a traditional *Capture-Recapture* method introduced in [Amstrup et al. 2005]. This book describes the first attempt by Pierre Simon Laplace to use the capture-recapture type of approach to estimate the size of a set of objects. In September of 1802, Laplace (1749-1827) used this method to estimate the size of the human population of France. He obtained census and live birth data from several sample communities across France. Also, all live births in France were recorded annually. He calculated the average of the total number of births reported per year (marked individuals, x) of the sampled communities and the total number of individuals (y). So the proportion that the marked individuals of the total individual to the sample communities is:

$$p = \frac{x}{y} \quad (1)$$

The individuals from the sampled communities could be considered sampled from all of France, then this proportion could be considered the proportion of births of the total population. Let N_p denotes the annual births of France, then the total population N could be estimated:

$$N = \frac{N_p}{p} \quad (2)$$

Theories and applications of capture-recapture share the basic concept, that is, **the ratio between the known and unknown**. Basically, the capture-recapture model captures a set of animals randomly, marks them and then releases them. After that, capture another sample and count the duplicates with the previous captures. By analyzing the overlap, the population of animals can be estimated. Some researchers derive their estimators from the *urn model* that considers the an urn containing N balls of different colors. Draw the balls randomly one at a time.

There are three basic schemes of returning balls. Most of the research consider all balls drawn are returned.

However, the above two models have an underlying assumption: each individual (like an animal and a ball) has equal probability to be captured. However, the capture of individual is not precisely random. Burnham and Overton [1979] first described the heterogeneity model that measured the varied capture probability. Burnham and Overton’s model measured the capture probability of i th individual on the j th occasion. This forms p_{ij} . They used a matrix to present the data. Each cell in the matrix is 1 or 0. 1 means the i th individual is captured on the j th occasion while 0 means is not captured. The frequency of an individual to be captured can be obtained. Chao in [Chao 1987] proposed a new estimator that was based on Burnham and Overton’s model. Chao extended their work and obtained the estimator that used an accumulative distribution function to estimate the overall frequency of individuals. Chao claims that his estimator produces reasonable results when the capture times increases. But Chao also remarked that if the average capture probability is large the proposed estimator fails to work.

Table I briefly summarizes the papers discussed in the background section.

Table I. Papers and contributions to the background knowledge of estimation methods

Year	Author(s)	Title of Paper	Main Contribution(s)
1979	Burnham and Overton	Robust estimation of population size when capture probabilities vary among animals	The authors described the heterogeneity model that measured the varied capture probability.
1987	Chao	Estimating the Population Size for Capture-Recapture Data with Unequal Catchability	This paper obtained the estimator that used an accumulative distribution function to estimate the overall frequency of individuals.
2005	Amstrup et al.	(book)Handbook of Capture-Recapture Analysis	This book introduces the basic methods of capture-recapture analysis approached and its apply models.

3. THE BASIC ESTIMATION METHOD

3.1 Query-based sampling

The tradition capture-recapture model and urn model use a random sampling technique and assume that the probability distribution of each individual satisfies a uniform distribution. However, deep web data sources do not provide public interfaces that allow arbitrarily access. Instead, only query interfaces are provided to allow public access. Hence, some estimation methods are based on issuing queries to access the data sources.

Obtaining a “good” random sample of a data source could be an ideal approach when the entire data source can not be completely downloaded. Two papers ([Si and Callan 2003] and [Karnatapu et al. 2004]) discuss the analysis of the random samples (resource description) of data source and then estimate the data source itself.

If the above method is not viable, researchers resorted to *query-based sampling* [Callan and Connell 2001]. Other than random sampling, when using query-based sampling, the probability distribution of each individual can be varied. For each capture (via issuing a query), each individual has unequal catchability. The bias of this sampling technique will be discussed in section 4.

3.2 Identifying documents

Query-based sampling requires each item returned to be identified. In the web context, there are many ways to identify an item returned by a query. Bharat and Broder [1998], and Lawrence and Giles [1998] used the URL to identify the web pages returned by Search Engines. When estimating the background databases of some commercial web sites, like Amazon, book ISBN or item ID could also be used to identify an item [Lu 2007][Ling et al. 2008].

3.3 Estimating relative size

To estimate the relative size of data sources, many methods are based on the following probability model: Consider sets A and B that have an intersection $A \cap B$:

- Let $P(A)$ denote the probability that an element belongs to set A .
- Let $P(A \cap B | A)$ denote the conditional probability that an element belongs to set $A \cap B$ simultaneously belonging to set A .

Then

$$P(A \cap B | A) = \frac{|A \cap B|}{|A|} \quad (3)$$

likewise,

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

Therefore

$$\frac{|A|}{|B|} = \frac{P(A \cap B | B)}{P(A \cap B | A)} \quad (5)$$

Bharat and Broder [1998] first proposed this method to estimate relative size of data sources that the relative size of sets A and B can be calculated by uniformly picking elements from sets A and B (**sampling**) and examining how many elements are identical (**checking**). These two steps, sampling and checking dominate the later estimation methods. Inevitably, all methods have the two basic estimation procedures.

Bharat and Broder [1998] conducted their experiment using four search engines: AltaVista, Excite, HotBot and Infoseek in two time periods with two conjunctive queries and two disjunctive queries. For each query to a search engine, a sample URL is selected from the top 100 results and sent to the other three search engines to check for containment. If the results from each search engines contain more than 10 URLs then this test will be discarded (as it is a strong query) and move on to the next one. They estimated HotBot had the largest coverage of the indexable web in June/July 1997 while AltaVista has the largest in November 1997.

In the same year, Lawrence and Giles [1998] claimed that the results of previous companies of search engine’s relative coverage has limited value because the search engine can often retrieve results that do not contain the terms for the query. Their method is similar to Bharat and Broder’s. But more precisely, they download the documents obtained by each query to check if each document contains the exact query terms. The documents that do not contain the query terms are removed. An important assumption that is neglected by previous research was proposed by the author: each search engine samples the web independently.

In this section, research on estimating the relative size of data sources have been discussed. Relative size estimation is based on the simple probability model that analyzes the overlap. The summary of papers and their contribution is given in Table II.

Table II. Papers and contributions on estimating relative size of data sources

Year	Author(s)	Title of Paper	Main Contribution(s)
1998	Bharat and Broder	A technique for measuring the relative size and overlap of public web search engines	The authors proposed basic method to estimate the search engines’ relative size. They also identified two main steps of estimation.
1998	Lawrence and Giles	Searching the World Wide web	The authors clarified several important assumptions applied in estimating relative size of data sources. Their method is based on the analysis of the overlap of two samples.

3.4 Estimating actual size

Now consider sets A and B that are randomly sampled from a data source that has N documents. There are three basic probability models to estimate N .

—Estimating the fraction p_a (p_b) of all documents N . Then,

$$N = \frac{|A|}{p_a} \quad (6)$$

or

$$N = \frac{|B|}{p_b} \quad (7)$$

—Sets A and B maybe too large or their sizes are unknown. Let A' and B' denote two random samples from A and B , respectively. Then N could be estimated by:

$$N = \frac{|B'|}{|A' \cap B'|} |A| \quad (8)$$

or

$$N = \frac{|A'|}{|A' \cap B'|} |B| \quad (9)$$

— N could also be estimated by:

$$N = \frac{|A| \times |B|}{|A \cap B|} \quad (10)$$

3.4.1 *Methods based on document IDs.* This section will provide an overview of estimation methods that only need document IDs in their methods. These methods consider a data collection as “a black box”. They do not download the documents returned by queries.

Bharat and Broder [1998] estimated the relative size of several search engines. Based on the the real size of one search engine, they inferred that the web should have 200 millions pages. Later, Lawrence and Giles [1998] estimated that the indexable web has 320 million pages (lower bound). Kishi et al. [2000] improved the method of Lawrence and Giles’s method [1998]. The new estimator obtained the N by averaging the results of equations 8 and 9. The query terms are collected from newspaper articles published in the Mainichi Shinbun between 1997 and 1998. They gave an estimate of 88 million pages in Japanese with a 95 percent confidence level interval of 1 million. This result is larger than other estimation results.

In 2005, Gulli and Signorini estimated that the indexable web is More than 11.5 Billion Pages [Gulli and Signorini, 2005]. Basically, they repeated Bharat and Broder’s method but built a query lexicon by indexing the whole DMOZ.com directory which has more than 4 million pages. All indexed pages resulted in a set of 2,190,702 terms. The terms were sorted by occurrence and divided in block of 20 terms. From each block, they picked up one term and 438,141 terms were obtained as a lexicon. Each query from the lexicon is “one-term” query. The author submitted the queries to Google, MSN, Yahoo! and Ask/Teoma. For each query, they selected at least one URL randomly from the first 100 results. In the sampling procedure, the authors the sent normalized the URLs to the interface provided by each search engines. The authors exploited the Helios meta-search engine to carry out sampling and checking procedures.

Liu et al. [2001] proposed a method that uses a basic *Capture-Recapture* methodology to estimate the size of a database. It was called the *Multi Capture-Recapture* method that repeats n times the process of randomly selecting a document from N and supposes Y of these n randomly selected documents are from m . Then the expected value of Y is:

$$E(Y) = n \times \frac{m}{N} \quad (11)$$

(m denotes the number of a random sample from the N documents) Take Y_0 as an approximation of Y to the $E(Y)$, then:

$$Y_0 \approx n \times \frac{m}{N} \quad (12)$$

The number of documents indexed by the search engine could be estimated by:

$$N \approx n \times \frac{m}{Y_0} \quad (13)$$

In their experiment, three text databases of different sizes were formed from TREC (Text REtrieval Conference) collections. For all three databases, the authors obtained a percentage error of less than 2.5%.

Bharat and Broder used paired terms sent to search engine in order to reduce the query bias [Bharat and Broder 1998]. Bolshakov and Galicia-Haro [2003] firstly analyzed the result of disjunctive queries to Google by taking two Spanish words (*que* and *de*). They found that the order of operands does not make much difference in the query results. However, it gives results differing by $\pm 17.3\%$ for the NOT operations. The authors inferred that the OR operation should not be used for estimation but no further proof was provided. The new method they proposed was built by a series of NOT operation. A few tens (functional word form) of words are selection from LEXSEP corpus of Spanish. These words are of the most frequency. They ordered the the forms according to their occurrence frequency in web pages. Starting from the first word, They ordered the words into the form that $word(1)$, $word(2)\neg word(1)$, $word(3)\neg word(2)\neg word(1)$, ... As a maximum, Google only accepts 10 words in a query. They summed the result of each form sent to query and count the unique documents the estimated number of pages Google has. They carried out experiments using two groups of 12 most frequent word forms of Spanish sent to Google. A tentative estimate result was 12,400,000 pages in Spanish using by Google.

Schumacher and Eschmeyer introduced *Capture Histories* method that uses a set of consecutive random samples with replacement [Schumacher and Eschmeyer 1943]. It considers the documents that have been captured prior to each step. In Shokouhi et al.'s experiment [Shokouhi et al. 2006], the *Multi Capture-Recapture* method and the *Capture-History* method produce close estimates after about 80 samples. However, when estimating the size of deep web data source, random sampling is not feasible. Shokouhi et al. proposed a correction to the *Capture Histories* method and *Multi Capture-Recapture* method to compensate the bias inherent in sampling via query-based sampling. The authors compensated the bias using training sets and applied regression analysis. The new methods are called *Capture Histories with Regression* and *Multi Capture-Recapture with Regression*. Their experimental results show that the *Multi Capture-Recapture with Regression* method “significantly outperforms the other algorithms in most cases and is robust to variations in collection size” [Shokouhi et al. 2006]. They claim that the capture history method is less expensive and more accurate than *Sample-Resample* method (will be discussed in 3.4.2), even using fewer probe queries.

The basic estimation model analyzes the overlap of two samples. Lu [2007] introduced a new term called *overlapping rate* (OR) which is the proportion of the number of accumulative documents returned by each query to the number of accumulative unique documents obtained. The author obtained the relation between the overlapping rate and percentage of corpus that returned documents covered by regression analysis. Using the Newsgroup corpus and Reuters data collections for training, the author obtained the overlapping law (theoretically proven in the paper) in English corpora:

$$P = 1 - OR^{-1.1} \quad (14)$$

New estimator:

$$\hat{n} = \frac{u}{P} = \frac{u}{1 - OR^{-1.1}} \quad (15)$$

The author proposed a new estimation method called the *OR method* which is an

iterative process. The input of the algorithm is terms and data source (corpus). The output is the size of the data source. The algorithm only needs document IDs when estimating instead of document content. In real applications, the document ID could be the URL of a web page returned by the search engines. The author considered that search engines usually return only a number of queries. In the experiment, the number of returned documents is set to 1000. When less than 1000 documents have been returned, all documents would be taken. Otherwise, it will randomly select 1000 documents. The author prepared six data collections and compared the estimation results obtained by the *OR method* with the *Capture Histories with Regression* method. After removing the overflowing queries, the author claims that the new method works better in dealing with ranked results of queries.

The table III summarized the authors and papers contributing to the methods that are based on document ID analysis.

Table III. Papers contributing to the methods that are based on document ID analysis.

Year	Author(s)	Title of Paper	Main Contribution(s)
1943	Schumacher and Eschmeyer	The estimation of fish populations in lakes and ponds	The authors introduced <i>Capture Histories</i> method.
1998	Lawrence and Giles	Searching the World Wide web	The authors clarified several important assumptions applied in estimating relative size of data sources. Their method is based on the analysis of the overlap of two samples.
2000	Kishi et al.	Estimating web properties by using search engines and random crawlers	The authors improved Lawrence and Giles's method [Lawrence and Giles, 1998].
2001	Liu et al.	Discovering the representative of a search engine	The authors proposed <i>multi capture recapture</i> method.
2003	Bolshakov and Galicia-Har	Can we correctly estimate the total number of pages in Google for a specific language?	The authors analyzed the impact on the result of disjunctive queries.
2005	Gulli and Signorini	The Indexable web is More than 11.5 Billion Pages	The authors analyzed the impact on the result of disjunctive queries.
2006	Shokouhi et al.	Capturing collection size for distributed non-cooperative retrieval	Proposed <i>multi capture recapture with regression</i> method and <i>capture histories with regression</i> method.
2007	Lu	Efficient Estimation of the Size of Text Deep web Data Source	The author proposed the OR method that analyzes the overlap between the unique documents and total documents after a set of consecutive queries.

3.4.2 *Methods based on document content analysis.* Some methods partially need to analyze the document content, obtaining terms and calculate the frequency of terms. Usually, the frequency of terms and the frequency of documents are two

important elements when compensating the bias (which we will discuss in section 4). Although the regression analysis approach is commonly used the the recent estimation methods, it needs training sets to obtain unknown variables. When applying the methods to estimate a new data collection, the bias could be varied.

In 2003, Wu et al. [2003] demonstrates the process of experiments that tried to find the relation between the number of words in a query with the unique documents that can be retrieved by a query. They first obtained the language model (a group of words with their frequencies) [Callan et al. 1999] by query-based sampling. And then they deleted the words which have the highest frequency in the language model. Next, they sent the low frequency words to query and collect the top N documents from all returned documents of queries. Finally, the number of unique document was computed. They tested the relationship between the number of sampling documents and the number of unique words that these sampling documents contain. Also, they carried out experiments to find the relationship between different number of words in a query and the percentage of unique documents out of total archive retrieved by this query.

Si and Callan [2003] pointed out that the cost of Liu et al.’s [2001] might be considered excessive. They proposed a new method called the *sample-resample* method. The new method assumes that a “good” (random) sample (N_{samp}) of a database (N), the **resource description** (a sample or subset of the database) has been provided. In their paper, they introduced the term *document frequency* w.r.t a query term: the number of documents that contain this query term. A denotes the event that a document sampled from the database contains term q_i . The probability of this event is:

$$P(A) = \frac{df_{q_i c_j - samp}}{N_{c_j - samp}} \quad (16)$$

B denotes the event that a document from the database contains term q_i . The probability of this event is:

$$P(A) = \frac{df_{q_i c_j}}{N_{c_j}} \quad (17)$$

(C_j denotes the database. $C_j - samp$ denotes the document sampled from the database when the resource description was created. N_{C_j} is the size of C_j . $N_{C_j - samp}$ is the size of $C_j - samp$. And q_i is the query term selected from the resource description for C_j . $df_{-q_i c_j}$ is the number of documents in C_j that contain q_i and $df_{-q_i c_j - samp}$ is the number of documents in $C_j - samp$ that contain q_i .) If the resource description is randomly created, the probability that a document sampled from the resource description contains a query term should be the same as the probability that a document sampled from the **database** contains the same query term. Then $P(A) \approx P(B)$. This bridges the “known” and “unknown”. The size of the database can now be estimated by:

$$\hat{N}_{C_j} = \frac{df_{q_i c_j} * N_{c_j - samp}}{df_{q_i c_j - samp}} \quad (18)$$

Si and Callan [2003] carried out experiments that compare the performance of the capture-recapture method and the sample-resample method with the new method

using two subsets of TREC123 collection. The new method has lower error rate.

In 2004, Karnatapu et al. [2004] proposed a new method, called the Independence Controlled Sample Size Estimation, which makes use of the resource description as well. This method is more complicated than Si and Callan's. One of the assumptions is the same as in [Si and Callan 2003]: The resource description is a good sample of the actual database. The other assumption requires that the database "provides information about the number of documents that match a given query" [Si and Callan 2003]. Differing from Si and Callan's method, Karnatapu et al. choose *independent* words to form pairs of words. To check the dependency of the words, the authors used two techniques to check the term independence: **independence criterion** and **chi-squared test**. Their method will first select a pair of independent terms (t_1 and t_2). Next, it will query the search engine using two terms in the pair, respectively (to obtain D_1 , D_2) and then conjunctively ($D_{1\cap 2}$). Then calculate an estimated size (est) by equation 10.

$$Est_i = \frac{D_1 * D_2}{D_{1\cap 2}} \quad (19)$$

Now, using the same scheme to query the resource description and calculate an estimated size (est'). Since the size of resource description is available, the ratio (called the *Correction Factor*) of estimated size of resource description and the size of resource description can be obtained.

$$CorrectionFactor(CF) = \frac{D_{R_1} * D_{R_2}}{D_{R_1 \cap R_2}} \quad (20)$$

Base on the assumption that the resource description is a "good" sample of the database, this ratio can be the bridge between the resource description and the database. As est is calculated, it multiplies by the Correct Factor to get the corrected estimation size of the database.

$$Est_{ic} = CF * Est_i \quad (21)$$

Continue this procedure until all pairs of terms have been used. Finally, sum all corrected estimation size and average it by the number of pairs. The final estimation result would be:

$$Est_f = \frac{\Sigma Est_{ic}}{n} \quad (22)$$

The authors claim that in their experiment, the *Independence Controlled Sample Size Estimation* method outperformed the *Sample-Resample* method.

There was a leap in the process of estimating collection size in 2006. In order to obtain a more accurate estimator, Bharat et al. introduced the terms *weight of a document* (w.r.t a query pool) and *weight of a query* (w.r.t a query pool). If a uniform random sample of documents in N can be obtained, a query pool can be built. The queries in the query pool are filtered from the random documents. *The weight of a document* is the reciprocal of the number of queries in the query pool A that this document d contains.

$$w_d^A = \frac{1}{d \cap A} \quad (23)$$

The weight of a query is the sum of all weights of documents that contain this query a .

$$w_a^A = \sum_{d \in D_A, a \in d} w_d^A \quad (24)$$

(D is the set of documents. For a query pool A , $D_A \subseteq D$ denotes a set of documents “such that every document in D_A contains at least one term in A ” [Karnatapu et al. 2004].) *The weight of a query pool* (w.r.t. the N documents) is the average of the weights of query in the query pool.

$$W_{A,D} = E_{a \sim A}[w_d^A] \quad (25)$$

The authors proved that the number of documents that the queries in query pool can obtain approximates to *The weight of a query pool* multiplied by the total number of all possible queries in the query pool:

$$W_{A,D} = \frac{|D_A|}{|A|} \quad (26)$$

This also implies that the query pool should have a closed boundary. In the experiment, the authors use all 5-digit numbers as a query pool. Let $p_A = \frac{|D_A|}{|D|}$, by equation 26, then

$$|D| = \frac{|A|}{p_A} \cdot W_{A,D} \quad (27)$$

The second estimator in this paper assumes that two uncorrelated query pools can be obtained. After using the query pool to obtain documents, use equation 10 to estimate the collection size. Let A and B denote two uncorrelated query pools, $D_{AB} \subseteq D$ denotes “the set of documents that contain at least one term in A and one term in B ” [Karnatapu et al. 2004], then

$$\frac{|D_{AB}|}{|D_A|} = W_{AB,D} = E_{a \in A} \left[\sum_{d \in D_{AB}, a \in d} w_d^A \right] \quad (28)$$

finally

$$|D| = \frac{|D_A| \cdot |D_B|}{D_{AB}} \quad (29)$$

In this estimation, the set of documents that contain at least one term in both query pools can be calculated similarly to the case if only one query pool exists. The authors use the TREC .gov collection to test the performance of two new estimators. In their experiment, the applied variance reduction techniques. For the first estimator, variance reduction techniques significantly improved the performs, 5,000 samples results 0.39% bias. However, the results of second estimator did not perform as accurately as the first estimator.

Although a lot of research has been done the on the metrics of search engines, designing automatic methods to measure the metrics is very challenging, especially in regards to accuracy and efficiency. Shokouhi et al. [2006] claim to have overcome the degree mismatch problem (the gap between the predicted document degree and actual one). They proposed two estimators using approximate weights which they

claim lead to nearly unbiased estimates. The first estimator is called the *Accurate Estimator* (AccEst) which “uses few search engine queries to probabilistically calculate exact document degrees” [Shokouhi et al. 2006]. The second estimator, *Efficient Estimator* (EffiEst), “predicts document degrees from the contents of documents alone” [Shokouhi et al. 2006]. They also observed that Broder et al.’s method implicitly applies Rao Blackwellization [Casella and Robert 1996] statistical tool for reducing estimation variance. The authors used Accurate Estimator to estimate the size of larger subsets of three real search engines, with or without duplicate elimination. The ODP collection split into two sets: a training set and a testing set. The estimator was evaluated to measure two metrics: The size of testing set (sum metric) and density of pages in the testing set about sports (average metric). The authors claim that the results of estimating corpus size shows that the two new estimators has small bias. Also, Rao-Blackwellization is effective in reducing the estimation variance. The estimators perform poorly in estimating the real search engines. And due to the impact of over-flowing queries, the estimator produced by Broder et al.’s estimator has larger bias than the new estimators.

After Shokouhi et al. used regression analysis approach to improve the accuracy of estimators, many researchers began to use this approach to obtain a more accurate estimator. Xu et al. [2007] proposed a new method, *Heterogeneous Capture* (HC) algorithm that models the capture probabilities of a document with logistic regression, and then estimates the collection size through conditional maximum likelihood. The covariates are the length len_i of document i , the static rank $rank_i$ of document i and the term frequency tf_{ij} of a query (term) j contained in the document i . Notice that the method can not collect the covariates for the documents never captured (because of the term frequency of a query in the document). The regression model is:

$$P_{ij} = \frac{\exp(\beta_0 + \beta_1 \cdot len_i + \beta_2 \cdot rank_i + \beta_3 \cdot tf_{ij})}{1 + \exp(\beta_0 + \beta_1 \cdot len_i + \beta_2 \cdot rank_i + \beta_3 \cdot tf_{ij})} \quad (30)$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ are unknown parameters. Border et al. [2006] measure the weight of a document. Xu et al. proposed a way to measure the probability of a document being captured. Let P_i denotes the probability of document i being captured at least once, then

$$P_i = 1 - \sum_{j=1}^k (1 - P_{ij}) \quad (31)$$

The estimator is:

$$\hat{N} = \sum_{i=1}^n \frac{1}{P_i} \quad (32)$$

Table IV summarizes the papers describing methods that are based on document analysis.

4. BIASES

Most of the estimation methods are query-based. Query-based sampling is not random hence it introduces bias. Bharat and Broder [1998] identified two major types of bias of query-based sampling: **query bias** and **ranking bias**.

Table IV. Papers that work on the methods that are based on document analysis

Year	Author(s)	Title of Paper	Main Contribution(s)
2003	Wu et al.	Experiments with Document Archive Size Detection	The authors tried to find the relation between number of words in a query with the unique documents can be retrieved by a query.
2003	Si and Clalan	Relevant document distribution estimation method for resource selection	The authors proposed the sample-resample method that utilizes the resource dscription .
2004	Karnatapu et al.	Estimating size of search engines in an uncooperative environment	The authors proposed the Independence Controlled Sample Size Estimation method.
2006	Broder et al.	Estimating corpus size via queries	The authors proposed two new methods by analyzing the document weight and query weight.
2006	Bar-Yossef and Gurevich	Random sampling from a search engine's index	Two new estimators using approximate weights are proposed. It leads to an unbiased estimate.
2007	Xu et al.	Estimating collection size with logistic regression	The authors introduce the <i>Heterogeneous Capture</i> algorithm.

4.1 Query Bias

Query bias is introduced by issuing queries to pick up documents. It is different from random sampling in that every document has an equal chance to be chosen. When firing a set of keywords to query the documents, “content rich” documents are more likely to be hit [Bharat and Broder 1998].

4.2 Rank Bias

Ranking bias is introduced by the search engines sorting of results, and only returning a certain amount of them [Bharat and Broder 1998].

5. OVERCOMING BIASES

5.1 Methods to Overcome Query Bias

Overcoming the query bias focuses on sending queries that make the distribution of catchability of documents close to a uniform distribution. Sending a term to query a document that is content-rich has higher chance to be captured. Some researchers use high frequency terms to capture documents. Then those content-rich documents could have a similar chance to be captured. However, in the case that the size of the documents are varied, the more small documents there are, the larger the bias will be. Other researchers use regression analysis approach and training set. However, when the collection changes, the bias will become unpredictable.

Bharat and Broder built a lexicon of roughly 400,000 words from 300,000 documents in Yahoo. Low frequency words were eliminated. They proposed two schemes to select random result pages from the query result pages returned by the search engines [Bharat and Broder 1998]:

- Choose keywords from the lexicon that have roughly the same frequencies to build disjunctive queries.

—Sort keywords by frequency, pair them with one coming from the upper threshold and the other coming from the lower threshold to build conjunctive query.

However, there is lack of theoretical proof of the schemes in this paper.

Lawrence and Giles [1998] fixed the maximum documents returned at 600 per query. They also claim that using a pair of larger search engines would have lower dependence when estimating relative size.

Shokouhi et al. improved the Capture-Recapture methods and Capture Histories that use regression analysis to compensate the bias [Shokouhi et al., 2006]. Lu inspected the relation of unique documents and total documents up to each query [Lu, 2007]. The cause of overlapping rate changes is the weight of query [Broder et al., 2006] varies.

5.2 Methods to Overcome Ranking Bias

When the return documents of each query is limited, different ranking algorithms generate different results. The existing methods all consider the “relevance ranking methods” (the documents that has high frequency of a term will be popular). However, when the ranking method changes, the estimators using document ID need to reform. This section will be limited as there are not many methods to overcome ranking bias.

Broder et al. [2006] identified those documents that have a lower chance to be captured, i.e. it needs more queries to hit the low frequency documents. They illustrated a generic method to identify the tail and truncate in the cases that the document frequency distribution satisfies either Monotone distributions or Power law distributions.

Some researchers use the regression analysis approach to overcome rank bias. Xu et al. approximate static rank by “the average position of the document in the returned lists across all the queries” [Xu et al. 2007].

6. OTHER ESTIMATION METHODS

Lawrence and Giles [2000] proposed a different method to estimate the total indexable web. The new method first random sampled IP addresses and web servers. And then crawled all pages from 2,500 random servers. The total number of pages could be estimated by

$$\hat{n} = \frac{\text{number of All Pages}}{2500} \times (\text{Number of All Possible Servers}) \quad (33)$$

Their experiment tested 4.3 billion possible IP addresses to estimate the total number of web servers. They estimated that public indexable web has 800 million pages.

Kishi et al. [2000] repeated the above method and gave an estimate for the number of web pages, 17 million in Japanese.

Although Bharat and Broder [1998], Liu et al. [2001], Si and Callan [2003], and Karnatapu et al. [2004] have proposed approaches to estimate the relative size of collections or search engines, they only use common search interface that provides one text field to enter the keywords. However, many deep web database query interfaces allow one to query on many attributes like “title” and “author”. No research that utilizes this feature to estimate the size of background databases. Ling et al. [2008] utilized different attributes that the interfaces of search engine provides. The

authors proposed a correlation matrix to record the dependencies of attributes an interface of a search engine provided. The lower dependency means that queries via two attributes will obtain fewer identical items. The authors stated that the dependency of two attributes can not reduce to 0, i.e. two attributes cannot be completely independent. Therefore, the sampling via two low dependent attributes is biased. The authors used two databases as training corpus. By regression analysis, the authors obtained the relation between *Dependency* and estimated bias ε .

Table V summarized the above papers and their contributions.

Table V. Other papers and contributions to estimating the size of data sources

Year	Author(s)	Title of Paper	Main Contribution(s)
2000	Lawrence and Giles	Accessibility of information on the web	The authors crawled all servers via all possible IP and random pages from sample IPs to estimate the total number of pages in web.
2008	Ling et al.	An attributes correlation that is based approach for estimating size of web databases	The authors analyzed the dependency of samples obtained by queries via two different query attributes provided by search engines.

7. CONCLUSION AND FUTURE WORK

In this survey, we have studied major developments in estimation methods of the size of deep web data sources. Many methods refers to the methods in biology that estimate the population size. During the period 1998 to 2003, starting from Bharat and Broder [1998], researchers are trying to determine many pages the search engines have indexed and further infer the size of indexable web. Early research focused on obtaining samples of search engines and used the tradition probability model to estimate the size. However, traditional probability model assumes samples can be randomly selected. Hence, the estimation of the size of search engines and the indexable web is heavily biased. The cause of this bias was identified later. A debate of whether Yahoo! is larger then Google aroused an interest to estimate the actual size of the data sources. After 2005, many researchers changed the direction to work on actual size estimation. There are two main trends in the development in estimating actual size: 1) consider the data source as a black box and 2) use the limited information obtained from/released by data sources. However, either way will not alleviate the bias, some methods compensate one but introduce another kind of bias. Some compensate for a set of particular training sets but they may not work well in estimating other data sources. From the literature on estimation methods, a low bias estimator could only work well in the the data sources it is “familiar” to.

Future work is needed in random sampling data sources. “It remains to develop a sampling technique which is applicable across a range of collections and which requires little prior knowledge of collection contents” [Thomas and Hawking 2007].

The future improvement in existing methods could be the combination of near-uniform query-based sampling approaches and utilize variance reduction techniques [Broder et al. 2006]. Randomly pair the words in low relevance (uncorrelated queries) so that a set of paired queries could capture as many unique documents as possible no matter what they are content-rich or not. The query interfaces of deep web data sources could provide many features. We infer that if it supports reverse ranking results then small size documents can also be captured. Since the query bias can not be completely removed, applying variance reduction techniques could further reduce the bias by analyze the samples returned by query.

8. ACKNOWLEDGEMENTS

I would like to thank Dr. Richard A. Frost for his instruction on conducting a literature review and survey, thanks also to Dr. Jianguo Lu for his suggestions on this topic and on the structure of this survey.

REFERENCES

- AMSTRUP, S. C., MCDONALD, T. L. AND MANLY, B. F. J. 2005. Handbook of Capture-Recapture Analysis. *Princeton University Press*, Oct, 2005.
- BAR-ILAN, J. 2006. Size of the web, search engine coverage and overlap – methodological issues. Unpublished, 2006.
- BAR-YOSSEF, Z., AND GUREVICH A. 2006. Random sampling from a search engine's index. *WWW 2006*. Edinburgh, Scotland, 367-376.
- BAR-YOSSEF, Z., AND GUREVICH A. 2007. Efficient search engine measurements. *WWW '07: Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 401-410.
- BERGMAN, M. 2001. The Deep Web: Surfacing Hidden Value. *JEP the Journal of Electronic Publishing*, 7(1), 2001.
- BHARAT, K. AND BRODER A. Z. 1998. A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. *Computer Networks*, 30(1-7), 379-388.
- BOLSHAKOV, I.A. AND GALICIA-HARO, S. N. 2003. Can we correctly estimate the total number of pages in Google for a specific language?. *Computational Linguistics and Intelligent Text Processing. 4th International Conference, CICLing 2003. Proceedings*, Mexico City, 415-419.
- BORDER, A. Z. 2000. Web measurements via random queries. *Presentation at the Workshop on Web Measurement, Metrics, and Mathematical Models (WWW10 Conference)*, 2000.
- Bourret, P. 1984. How to Estimate the Sizes of Domains. *Information Processing Letters*, 19(5), 237-243 (1984).
- BRIAND, L. C., EMAM, K. E. AND FREIMUT, B. G. 1998. A Comparison and Integration of Capture-Recapture Models and the Detection Profile Method. *In Proceedings of the Ninth international Symposium on Software Reliability Engineering*, Paderborn, Germany, 32-41.
- BRODER, A. Z., FONTOURA, M., JOSIFOVSKI, V., KUMAR, R., MOTWANI, R., NABAR, S. U., PANIGRAHY R., TOMKINS A. AND XU Y. 2006. Discovering the representative of a search engine. *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, Arlington, Virginia, USA, 594-603.
- BURNHAM, K. P. AND OVERTON, W. S. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5), 927-936.
- CALLAN, J., CONNELL M. AND DU A. 1999. Automatic discovery of language models for text databases. *SIGMOD Rec.*, 28(2), 479-490.
- CALLAN, J. AND CONNELL, M. 2001. Query-based sampling of text databases. *ACM Trans. Inf. Syst*, 19(2), 97-130, 2001.
- CASELLA, G. AND ROBERT, C. P. 1996. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81-94, 1996.
- CHAO, A., 1987. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, 43(4), 783-791.
- DOBRA, A. AND FIENBERG, S. E. 2004. How large is the World Wide Web?. *Web Dynamics*, 23-44, 2004.
- GULLI A. AND SIGNORINI A. 2005. The Indexable Web is More than 11.5 Billion Pages. *WWW 2005*, 902-903.
- HENZINGER, M. R., Heydon, A., Mitzenmacher, M., and Najork, M. 2000. On near-uniform URL sampling. *Computer Networks*, 295-308.

- HOLST, L. 1979. A Unified Approach to Limit Theorems for Urn Models. *Journal of Applied Probability*, 16(1), 154-162.
- JUHA M. A. 1990. Logistic Regression in Capture-Recapture Models. *Biometrics*, 46(3), 623-635.
- KISHI, N., OHMORI, T., SASAZUKA, S., KONDO, A., MIZUTANI, M. AND OGAWA, T. 2000. Estimating web properties by using search engines and random crawlers. *Proceedings of INET2000, The 10th Annual Internet Society Conference*, available at http://www.isoc.org/inet2000/cdproceedings/2a/2a_3.htm, 2000.
- KARNATAPU, S., RAMACHANDRAN K., WU, Z., SHAH, B., RAGHANAN V. V. AND RENTON, R. 2004. Estimating size of search engines in an uncooperative environment. *International Workshop on Web-based Support Systems*, Beijing, China, 81-87.
- LEE, S. AND CHAO, A. 1994. Estimating Population Size via Sample Coverage for Closed Capture-Recapture Models. *Biometrics*, 50(11), 88-97.
- LAWRENCE, S. AND LILES C. L. 1998. Searching the World Wide Web. *Science*, 280(5360), 98-100.
- LAWRENCE, S. AND LILES C. L. 2000. Accessibility of information on the Web. *Intelligence*, 11(1), 32-39.
- LIU, K., SANTOSO, A., YU, C. AND MENG, W. 2001. Discovering the representative of a search engine. *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, Atlanta, Georgia, USA, 577-579.
- LIU, K., YU, C., AND MENG, W. 2002. Discovering the representative of a search engine. *CIKM'02*, 652-654.
- LU J. 2007. Efficient Estimation of the Size of Text Deep Web Data Source. Unpublished, available at <http://cs.uwindsor.ca/~jlu/sigir2008.pdf>, 2007.
- LING, Y., MENG, X. AND LIU W. 2008. An attributes correlation based approach for estimating size of Web databases. *Journal of Software*, 19(2), 224-236.
- MOWSHOWITZ A. AND KAWAGUCHI A. 2005. Measuring Search Engine Bias. *Information Processing and Management*, 41(5), 1193-1205.
- MOTWANI, R., PANIGRAHY R., AND XU, Y. 2000. Estimating Sum by Weighted Sampling. *Lecture Notes in Computer Science*, vol.4596/2007, 53-64.
- Neumann, J. V. 1963. Various techniques used in connection with random digits. In *John von Neumann, Collected Works*, vol. V. Oxford, 1963.
- PRICE G. 2005. More on the total database size battle and Google whacking with Yahoo. available at blog.searchenginewatch.com/blog/050811-231448, 2005.
- RHODENIZER, D. T. D. 2004. Estimating the size and growth of the World Wide Web. M.Sc. diss., Acadia University (Canada).
- RUSMEVICHIENTONG, P., PENNOCK, D., LAWRENCE, S., AND GILES C. L. 2001. Methods for sampling pages uniformly from the World Wide Web. In *Proc. AAAI Symp. on Using Uncertainty within Computation*, 2001. 121-128.
- SI, L., AND CALLAN, J. 2003. Relevant document distribution estimation method for resource selection. *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, Toronto, Canada, 298-305.

- SCHUMACHER, F. X. AND ESCHMEYER, R. W. 1943. The estimation of fish populations in lakes and ponds. *Journal of the Tennessee Academy of Science*, 18:228–249.
- SHOKOUHI, M., ZOBEL J., SCHOLER F. AND TAHAGHOGHI, S. M. M. 2006. Capturing collection size for distributed non-cooperative retrieval. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA, 316-323.
- WU, S., GIBB, F. AND CRESTANI, F. 2003. Experiments with Document Archive Size Detection. *Advances in Information Retrieval, 25th European Conference on (IR) Research, (ECIR) 2003*, Pisa, Italy, 2003.
- XU, J., WU, S., AND LI, X. 2007. Estimating collection size with logistic regression. *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, 789-790.