Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

The challenges of producing statistics for the Web: sampling and automated data collection of webpage information in the Brazilian Web

Pedro Luis do Nascimento Silva[1], Emerson Gomes dos Santos[2],
Isabela Bertolini Coelho and Suzana Jaíze Alves da Silva[3]

Abstract

The Brazilian Network Information Center (NIC.br) has designed and carried out a pilot project to collect data from the Web in order to produce statistics about the webpages' characteristics. Studies on the characteristics and dimensions of the web require collecting and analyzing information from a dynamic and complex environment. The core idea was collecting data from a sample of webpages automatically by using software known as web crawler. The motivation for this paper is to disseminate the methods and results of this study as well as to show current developments related to sampling techniques in a dynamic environment.

Key Words: Automated data collection; Population size estimation; Sampling; Internet.

1. Introduction

1.1 Description

The Internet is probably the most sophisticated information and communication technology (ICT) currently available to society. Its structure and applications have many social, cultural, economic and political implications. The Web has become the most widely known application on the Internet and may be defined as the part of the Internet that can be accessed through browsers. Studies on the characteristics and dimensions of the Web require collecting and analyzing information from a dynamic and complex environment (CGI, 2010).

The Brazilian Network Information Center (NIC.br) has designed and carried out a pilot project to collect data from the Brazilian part of the Web in order to produce statistics about the characteristics of webpages such as size and age, language, types of objects embedded on the pages, technical data including protocols (IPv4, IPv6, HTML), and accessibility among others.

This pilot project was a first step towards establishing a methodology to collect the data in a dynamic environment without a frame. The core idea was collecting data from a sample of webpages automatically by using software known as a web crawler. Several methodological challenges related to sampling procedures were tackled in this project. The motivation for this paper is to disseminate the methods and results of this study as well as to show current developments related to sampling techniques in a dynamic environment.

[1] Pedro Luis do Nascimento Silva, IBGE - Escola Nacional de Ciências Estatísticas (ENCE), Rua André Cavalcanti, 106 - Bairro de Fátima - Rio de Janeiro RJ, Brasil, 20231-050 (pedronsilva@gmail.com);
[2] Emerson Gomes dos Santos, UNIFESP - Escola Paulista de Política, Economia e Negócios (EPPEN), Rua Angélica, 100 - Osasco - SP, Brasil, 06110-295 (emerson.gomes@unifesp.br);
[3]Isabela Bertolini Coelho (isabela@nic.br) and Suzana Jaíze Alves da Silva (suzana@nic.br), NIC.br - Núcleo de Informação e Coordenação do .Ponto BR, Avenida das Nações Unidas, 11.541 - Brooklin Novo - São Paulo SP, 04578-000.

2. Pilot Project Methodology

2.1 Overall Strategy

The ".br" part of the Web can be subdivided into smaller parts by considering the prefix part of the domain names. To illustrate, one could consider only web pages belonging to the ".gov.br" part of the Brazilian Web.

This subdivision is useful for stratification as well as for analytic purposes. Domains in the Generic Top-Level Domain (GTLD) ".gov.br" were the first to be surveyed in the context of the project. In the sequence, domains under the GTLD ".com.br" were surveyed using a probability sampling approach.

2.1 Web ".gov.br" Census Project

In order to conduct a first data collection aimed at developing and testing the data collection strategy and tools, the domains registered on behalf of Brazilian government under the GTLD ".gov.br" were chosen for taking a census. This GLTD was chosen due to its overall small size when compared to others domains (see Table 2.1-1). A census is a procedure of acquiring and recording information about every unit in a well specified population, and thus clearly depends on the definition of the population boundaries.

Table 2.1-1
Proportion (%) of Domains by Generic Top-Level Domain (GTLD) in the Brazilian Web

GTLD	Percent
.com.br	90.8%
.net.br	3.2%
.org.br	1.8%
.gov.br	0.1%
Others	4.1%

Source: NIC.br

The objectives of the Brazilian Web ".gov.br" census project were:

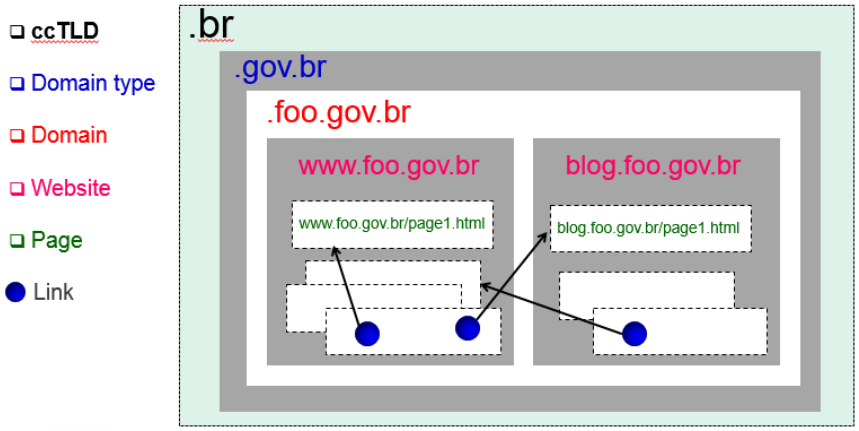
a) Measure the Brazilian government Web dimensions and characteristics; and
b) Provide indicators to describe websites under the gTLD ".gov.br".

The main indicators that this project aimed to produce related to characteristics of sites and pages, such as: Total Web size, Number of websites, Number of pages, Average age of pages, Languages used, Types of objects embedded, Technology used (open or proprietary), Compliance with W3C Standards, Web Accessibility Guidelines Proportions of web servers using IPv6, Host country (georeferenced IP), Synchronization to Coordinated Universal Time, and Average response time.

To carry out this study, it was necessary to define the key concepts of the units of analysis. The higher level units are "**Domains"**, which are referred to by a name under the ".gov.br" domain (figure 2.1-1). The second higher level units are "**Sites**", which are referred to by a name under a domain (figure 2.1-1). The lower level units of analysis considered were "**Pages**", which are referred to by complete names such as those indicated in figure 2.1-1. Finally, because the Web is a network interconnected through hypertext documents, the "**Links**" found within pages were used to navigate between pages, sites and domains.

This study was made possible over successive data collection of pages found from visiting an initial list of domains and sites called **seeds**. This means that the initial set of sites from which the search was conducted affects the final result, and finding the right initial set, as complete as possible, is a key step of the survey process. The initial list of domains identified as belonging to the GLTD ".gov.br" was provided by the registration authority for domain names in Brazil (Registro.br), responsible for the maintenance of the domains under the ".gov.br" by authorization from the Ministry of Planning and Budget. There were approximately 12 thousand domains in the initial set of domains, and for every domain in this list, a search was carried out automatically by software known as web crawlers (collectors).

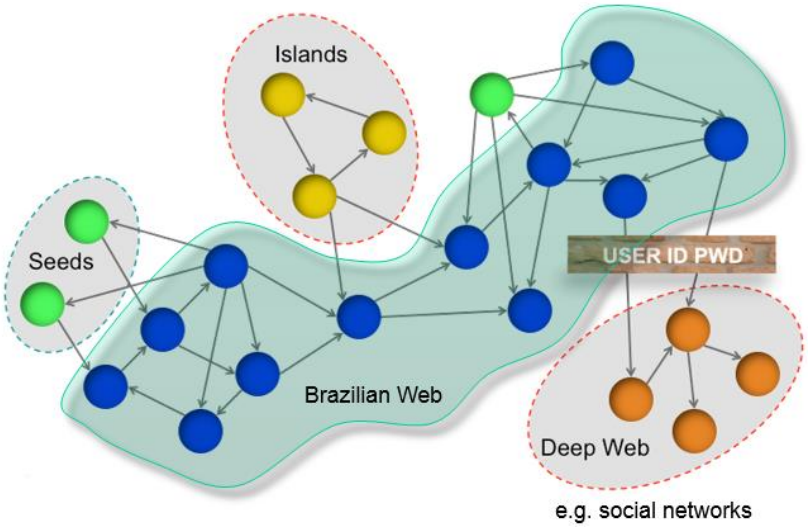
Figure 2.1-1
Units of analysis



Source: NIC.br

Although most of the Web is connected, limits of size and depth were established. There are "islands" of varying sizes with no connection to the rest of the network and there is the “deep web”, which is only accessible via user authentication. Therefore, the Web structure limits the technical ability to assess the real size and composition of what would be a “population of domains and technical objects” (figure 2.1-2).

Figure 2.1-2
Natural limits of Web



Source: NIC.br

Selected indicators of data collected from “.gov.br” are presented in Table 2.1-2. The overall size of the “.gov.br” by number of sites, number of pages and by size in Gigabytes were calculated from the census of “.gov.br” domains. Analysis of compliance of web pages with W3C standards was made considering the number of non-compliances found by the validating software. From all the sites with HTML pages collected, only 5% were fully compliant with the W3C standard.

Data collection (census) from all registered ".gov.br" domains, and from redirected domains if they were also ".gov.br" type, took about 3 weeks (≅12 thousand domains). Therefore, a census approach was considered infeasible for surveying the ".com.br" GTLD because data collection using the same tools as used for the ".gov.br" domains would last an estimated 11 years. Hence, the project adopted a sampling approach to survey the domains registered for Brazilian 'commercial' organizations / companies under the ".com.br" GTLD.

Table 2.1-2
Selected indicators for the ".gov.br" domains

Indicator	Estimate
Number of sites	11,856
Number of pages	6,331,256
Size in Gigabytes	169,7
Proportion W3C Compliant	5%

3. The Web Survey Project

3.1 Sampling Frame & Design

The Web is a dynamic environment, where rapid change takes place with domains, sites and pages being created and destroyed all the time. There is no readily available and up to date frame that can be used to sample sites and pages directly. There is however a list of all the **registered domains** which is kept by the Registro.br which regulates the Brazilian Web. This list is not an ideal frame for sampling because it may contain registered domains that may no longer exist on the Web (unreported 'deaths'), as well as because there may be active domains on the Web for which the corresponding record may have been removed from the list, for example because the organisation / person responsible may have failed to pay the corresponding registration fees.

Therefore a strategy was developed to use two sampling frames to extract samples of the Brazilian Web ".com.br" domains. First a frame (called A) was compiled with the names (of the form "*name*.com.br") of all domains under the".com.br" GTLD that were searched using the DNS (Domain Name System) server maintained by Registro.br over a period of 24 hours. In addition to the domain name, frame A contained information about the number of searches (NS) for each domain name, which we used as a measure of size to sample the domains. A total of 1,205,997 domains were included in frame A.

Frame A was then matched with the list containing 2,319,188 registered domains obtained from Registro.br, and the domains which were <u>not</u> searched during the specified 24 hour period were included in a separate frame (called B) where the only information available for each domain was the domain name. Therefore, Frame B contained only 1,113,191 domain names.

Given the scarcity of information about the population of sites and pages, a stratified single stage cluster sample design was adopted to survey the ".com.br" domains. The domains in Frame A were stratified by size using the stratification limits provided in Table 3.1-1. The stratum bounds were defined using the Geometric approach proposed by Gunning & Horgan (2004).

A total sample size of 3,000 domains was allocated to Frame A. After declaring strata 5 and 6 as certainty strata, the allocation of the remaining sample size to the strata 1-4 was carried out using Power Allocation with a power of ½.

Table 3.1-1
Sampling Design for com.br domains

Stratum h	Lower Limit	Upper Limit	Pop. Size N_h	Sample Size n_h
1	1	8	630,932	227
2	9	75	402,771	553
3	76	659	155,357	953
4	660	5,743	16,400	730
5	5,744	49,999	505	505
6	50,000	1,029,338	32	32
Total			1,205,997	3,000

A total sample size of 1,000 was allocated to Frame B. Data collection for this sample was attempted, but met with very limited success, and therefore results are not reported here, pending further analysis.

3.2 Data collection

Data collection of the sample of 3,000 domains from Frame A took about 3 months. Six domains were not found (non response?) during the data collection period. Within the sample domains, data were collected for 287,981 sites, out of which 20,476 sites were excluded because they resulted from redirection of sampled domains to non-sampled domains, most of which were blogs. Hence, the final sample of sites comprised 267,505 valid sites.

Some selected indicators (number of sites, number of pages and by size in Gigabytes) estimated using the ".com.br" sample from Frame A are presented in Table 3.2-1. Analysis of compliance of web pages with W3C standards was carried out considering all sites with HTML pages collected, and only 7% (s.e. 3.6%) were fully compliant with the W3C standard.

Table 3.2-1
Estimation for overall size of the com.br domains

Indicator	Estimate	Standard Error
Number of sites	1,621,242	161,855
Number of pages	289,803,146	44,781,861
Size in Petabytes (10^15 bytes)	14,014,479	3,876,370
Proportion W3C Compliant	7%	3.6%

4. Final comments

This pilot project was a first step towards establishing a methodology to collect the data in a dynamic environment without ideal frames. The core idea was collecting data from a sample of pages automatically by using software known as a web crawler. Several methodological challenges related to sampling procedures were tackled in this project. By studying the outcome of this pilot project we aim to develop improved procedures to survey the Brazilian Web in the future.

References

CGI.br (The Brazilian Internet Steering Committee), Dimensions and Characteristics of the Brazilian Web: a study of the gov.br. Available in: http://www.cgi.br/publicacoes/pesquisas/govbr/cgibr-nicbr-censoweb-govbr-2010.pdf

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons, Inc.

Gunning, P. & Horgan, J. M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. Survey Methodology 30: 159-166.

Särndal, C.E., Swensson, B. e Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer Verlag.
Thompson, S.K. (1999) Sampling. Wiley.