Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Défis de la production de statistiques pour le Web : échantillonnage et collecte des données automatisées de pages Web au Brésil

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

Résumé

Le Centre d'information de réseau brésilien (NIC.br) a conçu et mené un projet pilote pour recueillir des données sur le Web, afin de produire des statistiques concernant les caractéristiques des pages Web. Les études des caractéristiques et des dimensions du Web exigent la collecte et l'analyse de données dans un environnement dynamique et complexe. L'idée de base est de recueillir des données sur un échantillon de pages Web, automatiquement, en utilisant le logiciel appelé moteur de recherche Web. Le présent article vise à diffuser les méthodes et les résultats de cette étude, ainsi qu'à démontrer les progrès actuels liés aux techniques d'échantillonnage dans un environnement dynamique.

Mots clés : Collecte de données automatisée; estimation de la taille de la population; échantillonnage; Internet.

1. Introduction

1.1 Description

Internet est probablement la technologie de l'information et des communications (TIC) la plus raffinée actuellement disponible dans la société. Sa structure et ses applications comportent de nombreuses répercussions sociales, culturelles, économiques et politiques. Le Web est devenu l'application la plus connue sur Internet et peut être défini comme la partie d'Internet accessible au moyen de fureteurs. Les études des caractéristiques et des dimensions du Web exigent la collecte et l'analyse de données dans un environnement dynamique et complexe (CGI, 2010).

Le Centre d'information de réseau brésilien (NIC.br) a conçu et mené un projet pilote pour recueillir des données sur le Web brésilien, afin de produire des statistiques concernant les caractéristiques des pages Web, comme la taille et le relevé chronologique des pages, les langues, les types d'objets intégrés dans les pages, les données techniques, y compris les protocoles (IPv4, IPv6, HTML), et l'accessibilité, notamment.

Ce projet pilote est une première étape en vue de la mise en place d'une méthode pour recueillir des données dans un environnement dynamique, sans base de sondage. L'idée de base est de recueillir des données sur un échantillon de pages Web, automatiquement, en utilisant le logiciel appelé moteur de recherche Web. Plusieurs défis méthodologiques liés aux procédures d'échantillonnage ont été relevés dans le cadre de ce projet. Le but de cet

[1] Pedro Luis do Nascimento Silva, IBGE - Escola Nacional de Ciências Estatísticas (ENCE), Rua André Cavalcanti, 106 - Bairro de Fátima - Rio de Janeiro RJ, Brésil, 20231-050 (pedronsilva@gmail.com).
[2] Emerson Gomes dos Santos, UNIFESP - Escola Paulista de Política, Economia e Negócios (EPPEN), Rua Angélica, 100 - Osasco - SP, Brésil, 06110-295 (emerson.gomes@unifesp.br).
[3]Isabela Bertolini Coelho (isabela@nic.br) et 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.

article est de diffuser les méthodes et les résultats de cette étude, ainsi que de démontrer les progrès actuels liés aux techniques d'échantillonnage dans un environnement dynamique.

2. Méthodologie du projet pilote

2.1 Stratégie générale

La partie « .br » du Web peut être divisée en parties plus petites à partir du préfixe des noms de domaine. Par exemple, on pourrait tenir compte uniquement des pages Web appartenant aux domaines « .gov.br » du Web au Brésil.

Cette subdivision est utile pour la stratification ainsi que pour l'analyse. Les domaines appartenant au Domaine Générique de Premier Niveau (DGPN) « .gov.br » ont été les premiers à faire l'objet d'une enquête dans le contexte du projet. Dans la séquence, les domaines sous le DGPN « .com.br » ont été visés par une enquête reposant sur une approche d'échantillonnage probabiliste.

2.1 Projet de recensement Web à partir de « .gov.br »

Afin de procéder à une première collecte de données, en vue d'élaborer et de mettre à l'essai une stratégie et des outils de collecte des données, les domaines enregistrés au nom du gouvernement du Brésil sous le DGPN « .gov.br » ont été choisis pour la tenue d'un recensement. Ce DGPN a été choisi en raison de sa petite taille globalement comparativement à d'autres domaines (voir le tableau 2.1-1). Un recensement est une procédure d'acquisition et de consignation de données concernant chaque unité d'une population bien définie et, ainsi, dépend clairement de la définition des limites de la population.

Tableau 2.1-1
Proportion (%) des DGPN **sur le Web au Brésil**

Nom de domaine de premier niveau générique	*Pourcentage*
.com.br	90,8
.net.br	3,2
.org.br	1,8
.gov.br	0,1
Autres	4,1

Source : NIC.br

Les objectifs du projet de recensement Web à partir de « .gov.br » au Brésil étaient les suivants :

a) mesurer l'étendue et les caractéristiques du Web du gouvernement brésilien; et
b) fournir des indicateurs pour décrire les sites Web sous le DGPN « .gov.br ».

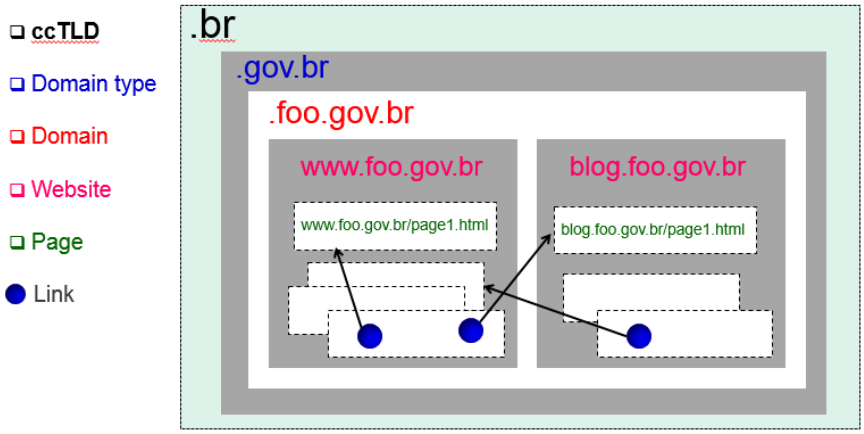
Les principaux indicateurs que ce projet visait à produire avaient trait aux caractéristiques de sites et de pages, par exemple : taille totale du Web, nombre de sites Web, nombre de pages, relevé chronologique des pages, langues utilisées, types d'objets intégrés, technologie utilisée (universelle ou privée), conformité aux normes W3C et aux directives sur l'accessibilité du Web, proportions des serveurs Web utilisant le protocole IPv6, pays hôte (IP géoréférencé), synchronisation avec le temps universel coordonné et délai de réponse moyen.

Pour mener cette étude, il était nécessaire de définir les concepts clés des unités d'analyse. Les unités de niveau plus élevé sont les « **domaines** », qui sont désignés au moyen d'un nom sous le domaine « .gov.br » (figure 2.1-1). Les unités du deuxième niveau en importance sont les « **sites** », qui sont désignés au moyen d'un nom sous un domaine (figure 2.1-1). Les unités d'analyse de niveau inférieur prises en compte étaient les « **pages** », qui sont désignées par

des noms complets, comme ceux indiqués dans la figure 2.1-1. Enfin, étant donné que le Web est un réseau interrelié, grâce à des documents hypertextes, les « **liens** » se trouvant dans les pages ont servi à naviguer entre les pages, les sites et les domaines.

Cette étude a été rendue possible grâce à la collecte successive de données de pages trouvées après une visite d'une liste initiale de domaines et de sites appelés **germes**. Cela signifie que l'ensemble initial de sites à partir desquels la recherche a été effectuée a des répercussions sur les résultats finaux, et que la détermination de l'ensemble initial approprié, le plus complet possible est une étape clé du processus d'enquête. La liste initiale de domaines identifiés comme appartenant au DGPN « .gov.br » a été fournie par les responsables de l'enregistrement des noms de domaine au Brésil (Registro.br), qui sont chargés de la mise à jour des domaines sous « .gov.br », par suite d'une autorisation du ministère de la Planification et du Budget. On comptait environ 12 000 domaines dans l'ensemble initial de domaines, et pour chaque domaine de cette liste, une recherche automatique a été effectuée au moyen d'un logiciel appelé moteur de recherche Web (collecteurs).

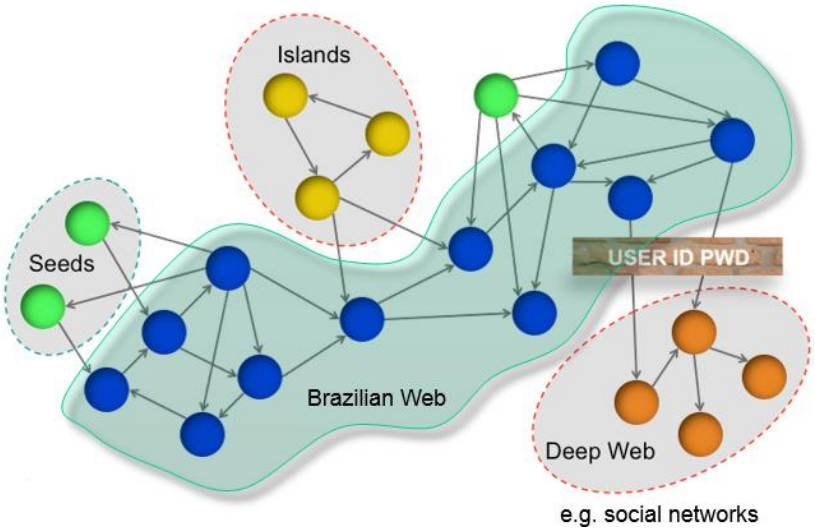
Figure 2.1-1
Unités d'analyse



Source : NIC.br

Même si la majeure partie du Web est reliée, des limites de taille et de profondeur ont été établies. Il y a des « îlots » de diverses tailles qui ne sont pas reliées au reste du réseau, et il y a le « Web caché », qui est uniquement accessible après une authentification de l'utilisateur. Par conséquent, la structure du Web limite la capacité technique à évaluer la taille réelle et la composition de ce qui serait une « population de domaines et d'objets techniques » (figure 2.1-2).

Figure 2.1-2
Limites naturelles du Web



Source : NIC.br

Certains indicateurs des données recueillies à partir de « .gov.br » sont présentés dans le tableau 2.1-2. La taille globale de « .gov.br », selon le nombre de sites, le nombre de pages et la taille en gigaoctets, a été calculée à partir du recensement des domaines « .gov.br ». L'analyse de la conformité des pages Web aux normes W3C a été effectuée en tenant compte du nombre de cas de non-conformité trouvés par le logiciel de validation. Parmi tous les sites dont les pages HTML ont été recueillies, seulement 5 % étaient entièrement conformes à la norme W3C.

La collecte des données (recensement) à partir de tous les domaines « .gov.br » enregistrés et à partir des domaines redirigés, s'ils étaient aussi du type « .gov.br », a pris environ trois semaines (≅12 000 domaines). Par conséquent, une approche de recensement a été jugée non viable pour une enquête sur le DGPN « .com.br », parce que la collecte des données au moyen des mêmes outils que pour les domaines « .gov.br » prendrait environ 11 ans. Ainsi, dans le cadre du projet, on a adopté une approche d'échantillonnage pour faire l'enquête sur les domaines enregistrés au nom de sociétés/entreprises « commerciales » du Brésil sous le DGPN « .com.br ».

Tableau 2.1-2
Certains indicateurs des domaines « .gov.br »

Indicateur	Estimation
Nombre de sites	11 856
Nombre de pages	6 331 256
Taille en gigaoctets	169,7
Proportion de conformité à W3C	5 %

3. Projet d'enquête Web

3.1 Base d'échantillonnage et conception

Le Web est un environnement dynamique dans lequel ont lieu des changements rapides; des domaines, des sites et des pages étant sans cesse créés et détruits. Il n'y a pas de base de sondage facilement disponible et à jour qui peut être utilisée pour échantillonner les sites et les pages directement. Il existe toutefois une liste de tous les **domaines**

enregistrés, qui est conservée par Registro.br, organisme qui réglemente le Web au Brésil. Cette liste ne représente pas une base idéale pour l'échantillonnage parce qu'elle peut contenir des domaines enregistrés qui n'existent peut-être plus dans le Web (« disparitions » non déclarées), et aussi parce qu'il peut y avoir des domaines actifs sur le Web dont l'enregistrement peut avoir été supprimé de la liste, par exemple, parce que l'organisation/la personne responsable a omis de payer les frais d'enregistrement correspondants.

Par conséquent, une stratégie a été élaborée pour utiliser deux bases d'échantillonnage, en vue d'extraire des échantillons des domaines « .com.br » du Web au Brésil. Tout d'abord, une base (appelée A) a été compilée au moyen des noms (de la forme « *name*.com.br ») de tous les domaines sous le DGPN « .com.br », qui a été recherché au moyen du serveur de nom de domaine (DNS) de Registro.br, sur une période de 24 heures. Outre le nom de domaine, la base A comprenait des renseignements concernant le nombre de recherches pour chaque nom de domaine, que nous avons utilisé comme mesure de la taille d'échantillonnage des domaines. Au total, 1 205 997 domaines ont été inclus dans la base A.

La base A a par la suite été appariée à la liste comprenant 2 319 188 domaines enregistrés, obtenus à partir de Registro.br, et les domaines qui n'ont pas été recherchés pendant la période déterminée de 24 heures ont été inclus dans une base distincte (appelée B), dans laquelle la seule information disponible pour chaque domaine était le nom du domaine. Par conséquent, la base B comprenait uniquement 1 113 191 noms de domaine.

Compte tenu de la rareté de l'information concernant la population des sites et des pages, un échantillon en grappes stratifié à un degré a été adopté pour faire l'enquête sur les domaines « .com.br ». Les domaines de la base A ont été stratifiés selon la taille, au moyen des limites de stratification fournies dans le tableau 3.1-1. Les limites de la strate ont été définies au moyen de l'approche géométrique proposée par Gunning et Horgan (2004).

Une taille d'échantillon totale de 3 000 domaines a été tirée de la base A. Après déclaration des strates 5 et 6 comme strates de certitude, l'affectation du reste de la taille de l'échantillon aux strates 1 à 4 a été effectuée au moyen d'une répartition par puissance avec puissance de ½.

Tableau 3.1-1
Plan d'échantillonnage pour les domaines com.br

Strate h	Limite inférieure	Limite supérieure	Taille de la population N_h	Taille de l'échantillon 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

Une taille d'échantillon totale de 1 000 a été tirée de la base B. La collecte des données pour cet échantillon a été tentée, mais a obtenu un succès très limité et, par conséquent, les résultats ne sont pas présentés ici, en attendant une analyse plus poussée.

3.2 Collecte des données

La collecte des données de l'échantillon de 3 000 domaines de la base A a pris environ trois mois. Six domaines n'ont pas été trouvés (non-réponse?) pendant la période de collecte des données. À l'intérieur des domaines de l'échantillon, des données ont été recueillies pour 287 981 sites, en excluant 20 476 sites parce qu'ils découlaient de domaines échantillonnés redirigés vers des domaines non échantillonnés, dont la plupart étaient des blogues. Ainsi, l'échantillon final de sites comprenait 267 505 sites valides.

Certains indicateurs (nombre de sites, nombre de pages et taille en gigaoctets) estimés à partir de l'échantillon « .com.br » de la base A sont présentés dans le tableau 3.2-1. L'analyse de la conformité des pages Web aux normes W3C a été effectuée en tenant compte de tous les sites ayant des pages HTML recueillies, et seulement 7 % (erreur type de 3,6 %) étaient entièrement conformes à la norme W3C.

Tableau 3.2-1
Estimation de la taille globale des domaines com.br

Indicateur	Estimation	Erreur type
Nombre de sites	1 621 242	161 855
Nombre de pages	289 803 146	44 781 861
Taille en pétaoctets (10^15 octets)	14 014 479	3 876 370
Proportion de conformité à W3C	7 %	3,6 %

4. Commentaires finaux

Ce projet pilote est une première étape en vue de la mise en place d'une méthode pour recueillir des données dans un environnement dynamique, sans base de sondage idéal. L'idée de base était de recueillir des données pour un échantillon de pages Web, automatiquement, en utilisant le logiciel appelé moteur de recherche Web. Plusieurs défis méthodologiques liés aux procédures d'échantillonnage ont été relevés dans le cadre de ce projet. Grâce à l'étude des résultats de ce projet pilote, nous visons l'élaboration de procédures améliorées pour faire enquête sur le Web au Brésil à l'avenir.

Bibliographie

CGI.br (Comité directeur de l'Internet brésilien). Dimensions et caractéristiques du Web brésilien : une étude de gov.br. Disponible à : http://www.cgi.br/publicacoes/pesquisas/govbr/cgibr-nicbr-censoweb-govbr-2010.pdf.

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

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

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