Sampling the National Deep Web

Denis Shestakov

Department of Media Technology,
Aalto University, Espoo, 02150 Finland
denis.shestakov@aalto.fi

Abstract. A huge portion of today's Web consists of web pages filled with information from myriads of online databases. This part of the Web, known as the deep Web, is to date relatively unexplored and even major characteristics such as number of searchable databases on the Web or databases' subject distribution are somewhat disputable. In this paper, we revisit a problem of deep Web characterization: how to estimate the total number of online databases on the Web? We propose the Host-IP clustering sampling method to address the drawbacks of existing approaches for deep Web characterization and report our findings based on the survey of Russian Web. Obtained estimates together with a proposed sampling technique could be useful for further studies to handle data in the deep Web.

Keywords: deep Web, web databases, web characterization, DNS load balancing, virtual hosting, Host-IP clustering, random sampling, national web domain.

1 Introduction

The deep Web, the huge part of the Web consisting of web pages accessible via web search forms (or search interfaces), is poorly crawled and thus invisible to current-day web search engines [13]. Though the problems with crawling dynamic web content hidden behind form-based search interfaces were evident as early as 2000 [6], the deep Web is still not adequately characterized and its key parameters (e.g., the total number of deep web sites and web databases, the overall size of the deep Web, the coverage of the deep Web by conventional search engines, etc.) can only be guessed.

Until now only a few efforts on the deep Web characterization have been done [6,9,16] and, more than that, one of these works is a white paper, with all findings obtained by using proprietary methods. Two other surveys, while being methodologically sound and reproducible, have inherent limitations due to the random sampling of IP addresses (rsIP for short) approach used in them. The most serious drawback of the rsIP method is the neglect of *virtual hosting*, i.e., a common practice of sharing one web server by multiple web sites. One of the largest European web hosting companies, OVH, with its 65,000 servers hosting over 7,500,000 sites[1] can be an example of actual ratios of web sites to servers

[1] See http://www.ovh.co.uk/aboutus/ovh_figures.xml (retrieved in May 2010).

A. Hameurlain et al. (Eds.): DEXA 2011, Part I, LNCS 6860, pp. 331–340, 2011.

on the Web. The rsIP, however, usually detects from only one to three web sites per a server and ignores the rest, regardless of the actual number of sites on the server. In the context of deep Web characterization, ignoring the virtual hosting factor means that a portion of web sites (hosted on servers with sampled IP addresses) is overlooked, making the estimates obtained in the abovementioned surveys seriously biased.

Our contributions. We propose a novel method for sampling the deep Web, the Host-IP cluster sampling technique. Our approach is based on the idea of clustering hosts sharing the same IP addresses and analyzing "neighbors by IP" hosts together. Usage of host-IP mapping data allows us to address drawbacks of the rsIP used in previous deep Web surveys, specifically to take into account the virtual hosting factor. While we designed the proposed method for the survey of deep web resources on a national segment of the Web, it could also be applied to more general characterization studies of the entire Web.

Experimental results. To validate our technique, we applied the proposed approach to study the Russian segment of the deep Web. We used over 670,000 hostnames to generate around 80,000 groups of hosts which were then sampled, crawled and examined for the presence of search forms. Based on the results of sample analysis, we obtained statistically significant estimates for the total number of deep web sites and web databases in the Russian Web. We also compared our technique with the rsIP method and observed that the rsIP applied to the same data would result in substantial underestimations, e.g., would detect less than a third of deep web resources. Additionally, we demonstrated the magnitude of virtual hosting by reviewing previous web surveys (Section 2) and by resolving over 670,000 hostnames to their IP addresses (Section 5).

The next section gives a background on methods to characterize the deep Web. In Sections 3 and 4 we present our approach, the Host-IP cluster sampling technique. In Section 5 we report the results of the survey and compare the proposed method with the rsIP method. We then outline prior works and briefly discuss some of our findings in Section 6. Finally, Section 7 concludes the paper.

2 Background: Deep Web Characterization

Existing attempts to characterize the deep Web [6,9,16] are based on two methods originally applied to general Web surveys: namely, *overlap analysis* [7] and *random sampling of IP addresses* [12]. The first technique involves pairwise comparisons of listings of deep web sites, where the overlap between each two sources is used to estimate the size of the deep Web (specifically, total number of deep web sites) [6]. The critical requirement to listings be independent from one another is unfeasible in practice; making the estimates produced by overlap analysis seriously biased. Additionally, the method is generally non-reproducible.

Unlike the overlap analysis the second technique, the random sampling of IP addresses technique (rsIP), is easily reproducible and requires no pre-built listings. The rsIP estimates the total number of deep web sites by analyzing a

sample of unique IP (Internet Protocol) addresses randomly generated from the entire space of valid IPs and extrapolating the findings to the Web at large. Since the entire IP space is of finite size and every web site is hosted on one or several web servers, each with an IP address[2], analyzing an IP sample of adequate size can provide reliable estimates for the characteristics of the Web in question. In [9], one million unique randomly-selected IP addresses were scanned for active web servers by making an HTTP connection to each IP. Detected web servers were exhaustively crawled and those hosting *deep web sites* (defined as web sites with search interfaces, or search forms, that allow a user to search in underlying databases) were identified and counted.

Unfortunately the rsIP approach has several limitations. The most serious drawback is ignoring virtual hosting, i.e., the fact that multiple web sites can share the same IP address. This leads to ignoring a certain number of sites, some of which are apparently deep web sites. The reverse IP procedure (applied to obtain a hostname based on an IP address) typically identifies one or two web sites hosted on a given IP, while hosting a lot of sites on the same IP is a common practice. As an example, according to [1], 4.4 millions of IP addresses hosted almost 50 millions of hosts in April 2004 or, on the level of national domains, 640 thousands of second-level domain names in .RU and .SU zones resolved to 68 thousands of IPs in March 2007 (see the survey at `http://www.rukv.ru/runet-2007.html`).

Another factor overlooked by the rsIP method is *DNS load balancing*, i.e., the assignment of multiple IP addresses to a single web site. For instance, Russian news site `newsru.com` mapped to three[3] IPs is three times more likely to appear in a sample of random IPs than a site with one assigned IP. Since the DNS load balancing is the most beneficial for popular and highly trafficked web sites we expect that the bias caused by the load balancing is less than the bias due to the virtual hosting. Indeed, according to the SecuritySpace's survey as of April 2004, only 4.7% of hosts had their names resolved to multiple IP addresses [2], while more than 90% of hosts shared the same IP with others [1].

To summarize, the virtual hosting cannot be ignored in any IP-based sampling survey. Next we propose a new sampling strategy to address these challenges.

3 Our Approach

Real-world web sites are hosted on several web servers, share their web servers with other sites, and are often accessible via multiple hostnames. Neglecting these issues makes estimates produced by IP-based or host-based sampling seriously biased.

The clue to a better sampling strategy lies in the fact that hostname aliases for a given web site are frequently mapped to the same IP address. In this way, given

[2] An IP address is not a unique identifier for a web server as a single server may use multiple IPs and, conversely, several servers can answer for the same IP.

[3] Here and hereafter if not otherwise indicated, resolved in May 2010.

a hostname resolved to some IP address, we can identify other hostnames potentially pointing to the same web content by checking other hostnames mapped to this IP. It is interesting to see here a strong resemblance to the virtual hosting problem, where all hosts sharing a given IP address have to be found. Assuming a large listing of hosts is available, we can acquire the knowledge about which hosts mapped to which IPs by resolving all hostnames in the listing to their corresponding IP addresses. Technically, such massive resolving of available hosts to their IPs is essentially a process of clustering hosts into groups, each including hosts sharing the same IP address. Grouping hosts with the same IPs together is quite natural because it is exactly what happens on the Web, where a web server serves requests only to those hosts that are mapped to a server's IP. Once the overall list of hosts is clustered by IPs we can apply a cluster sampling strategy, where an IP address is a *primary sampling unit* consisting of a cluster of *secondary sampling units*, hosts.

Our Host-IP approach addressing all the drawbacks described in the previous section consists of the following major steps:

- *Resolving, clustering and sampling:* resolve a large number of hosts relating to a studied web segment to their IP addresses, group hosts based on their IPs, and generate a sample of random IP addresses from a list of all resolved IPs.
- *Crawling:* for each sampled IP crawl hosts sharing a sampled IP to a predefined depth. While crawling new hosts (which are not in the initial main list) may be found: those mapped to a sampled IP are to be analyzed, others are analyzed if certain conditions met (see Section 4.2).
- *Deep web site identification:* Analyze all pages retrieved during the crawling step and detect those with search interfaces to databases.

One can notice the principal difference between a sample unit of the rsIP method and a sample unit of the Host-IP approach. While all sampling units (IPs) in the rsIP are fully identical among each other, a sample of units in the Host-IP is heterogeneous. Indeed, in the Host-IP method there is an associated cluster of hosts for every sampled IP, and these clusters vary in size. Therefore, it could be useful to stratify, i.e., divide the resolved list of IP addresses into several non-overlapping parts (strata) and then deal with each part independently. The reasoning behind such separation is a reasonable assumption that deep web sites are more likely to be found within groups of hosts of certain sizes. If so, it might be beneficial to study groups with a few hosts separately from groups including hundreds of hosts. Another support for stratification is in the fact that IP addresses referred to a large number of hosts are good indicators of server hosting spam web sites [10] and, hence, deep web sites are less likely to be found among such IPs. Yet another reason is to actually verify whether deep web sites are running on servers hosting only a few sites. Anyhow, we note that the stratification itself is only a supplemental step and can easily be omitted.

4 Host-IP Cluster Sampling

In this section we give the schematic description of the Host-IP cluster sampling technique. The detailed description can be found in [15].

4.1 Dataset Preparation and Sampling

The steps for dataset preparation, clustering and sampling are as follows:

1. Obtain a set of unique host-IP pairs by resolving an available set of hosts to their IP addresses.
2. Remove host-IP pairs with invalid IP addresses.
3. Divide a set of host-IP pairs into two subsets S and S^*: S, which is used for clustering at step 4, has exactly one host-IP pair for each host and S^* includes all remaining pairs. Such separation allows us to avoid dealing with the DNS load balancing factor at the next clustering and sampling steps.
4. Group host-IP pairs in S by IPs and (optionally) stratify groups obtained by their sizes. Host-IP pairs with the same IP form a group. As a result, we obtain N groups, where N is the number of unique IP addresses among the pairs in S. Denote a set of all unique IPs in S as I and a set of all hosts in S as H. The number of pairs in a group (or in other words the number of hosts sharing a given IP) defines the group size and can be used as the stratification parameter.
5. Randomly select n IPs from I or, if stratified, for each stratum randomly select n_k IPs from I_k.

Now obtained sample (or, if stratified, samples) of IPs can be processed according to the crawling strategy presented next.

4.2 Crawling Strategy

Each IP in a given sample is processed independently from other IPs in this sample. The steps of the algorithm to crawl hosts (secondary sampling units) associated with a sampled IP are:

1. For each sampled IP ip, $ip \in I(I_k)$ extract from $S(S_k)$ a set of hosts H_{ip} sharing ip.
2. Each host in H_{ip} is crawled to a predefined depth. Crawling (i.e., following links) is done selectively: a link leading to a host that belongs to $H \setminus H_{ip}$ is not followed to not violate the sampling procedure. All other links are followed. Since it is expected that H has no full coverage of the studied web segment, we pay special attention to hosts out of H. So, while crawling hosts in H_{ip}, we add all unknown hosts to a set of hosts H^u_{ip}, i.e., $H^u_{ip} \cap H = \emptyset$.
3. After completion of crawling H_{ip} we proceed to crawl all hosts in H^u_{ip} to a predefined depth. Similar to the previous step following links is selective. A link to a host h' is followed only if $h' \notin H \setminus H_{ip}$ and (h' is in $H^u_{ip} \cup H_{ip}$ or h' is a subdomain of one of the hosts in $H^u_{ip} \cup H_{ip}$ or h' is resolved to ip). Unlike step 2 unknown hosts are not collected anymore.

After last IP in the sample is crawled, all pages retrieved are inspected according to the identification process revealed in the next section.

4.3 Deep Web Site Identification

All pages retrieved during the crawling step are analyzed for the presence of search forms (interfaces to web databases). In order to consider just unique search forms, pages with duplicated forms are removed. At the start, we exclude pages without web forms and then, based on the methodology described in [14], pages with non-searchable forms (i.e., forms that are not interfaces to databases such as forms for site search, navigation, login, registration, subscription, polling, posting, etc.). Identified pages with search interfaces are then grouped by their web sites. Web sites with two or more search forms are additionally studied to determine how many web databases are actually accessible via a particular site.

4.4 Estimates for Total Number of Deep Web Sites and Databases

Let N_k and n_k represent the total and sampled numbers of IP addresses of the k-th stratum correspondingly and h_{ki} the number of hosts on the i-th IP ($1 \leq i \leq N_k$) of the k-th stratum. The total number of hosts in stratum k is $H_k = \sum_{i=1}^{N_k} h_{ki}$, and the number of analyzed (sampled) hosts in stratum k is $h_k = \sum_{i=1}^{n_k} h_{ki}$. Let s_{ki} (d_{ki}) denote the number of deep web sites (databases) detected among the hosts on the i-th IP of stratum k. $s_{ki}(d_{ki}) = 0$ if no deep web sites (databases) are detected, $s_{ki}(d_{ki}) > 0$ otherwise. Then, according to Chapter 12, p.116 of [17], the estimate for the total number of deep web sites (databases) in the k-th stratum $\widehat{S_k}$ ($\widehat{D_k}$) is: $\widehat{S_k} = \frac{H_k}{h_k} \sum_{i=1}^{n_k} s_{ki}$ $\left(\widehat{D_k} = \frac{H_k}{h_k} \sum_{i=1}^{n_k} d_{ki}\right)$.

The estimator of the variance of $\widehat{S_k}$ is given by

$$\widehat{var}(\widehat{S_k}) = \frac{n_k(N_k - n_k)H_k^2}{N_k(n_k - 1)h_k^2} \sum_{i=1}^{n_k} (s_{ki} - h_{ki}\frac{\sum_{i=1}^{n_k} s_{ki}}{h_k})^2 \quad (1)$$

The estimator of the variance of $\widehat{D_k}$ is identical to (1) except that all s_{ki} in the formula should be replaced with d_{ki}. The approximate 95% confidence interval for the total number of deep web sites (databases) in the k-th stratum is provided by $\widehat{S_k} \pm t\sqrt{\widehat{var}(\widehat{S_k})}$ $\left(\widehat{D_k} \pm t\sqrt{\widehat{var}(\widehat{D_k})}\right)$, where t is the upper 0.025 point of Student's t distribution with $n_k - 1$ degrees of freedom. Finally, the approximate 95% confidence interval for the total number of deep web sites (databases) in all strata is

$$\sum_{k=1}^{L} \widehat{S_k} \pm t\sqrt{\sum_{k=1}^{L} \widehat{var}(\widehat{S_k})} \quad \left(\sum_{k=1}^{L} \widehat{D_k} \pm t\sqrt{\sum_{k=1}^{L} \widehat{var}(\widehat{D_k})}\right), \quad (2)$$

where L is the number of strata.

5 Experiments

For our experiments conducted in September 2006 we used two lists of hostnames from datasets "Hostgraph" and "RU-hosts". We merged them into one list of

Table 1. Number of IP addresses and hosts (total and sampled) in each stratum

Strata	Num of IPs:	Num of sampled IPs:	Num of hosts:	Num of sampled hosts:
Stratum 1 (S1)	71486	964	112755	1490
Stratum 2 (S2)	5390	100	86829	1584
Stratum 3 (S3)	1860	11	472474	3163

unique hostnames. Next, following the methodology described in Section 4.1, we built the dataset for our survey. We resulted in 717,240 host-IP pairs formed by 672,058 unique hosts and 79,679 unique IP addresses. These numbers specifically show us that DNS load balancing has a modest influence – only 5.4% (36,349) of hosts are mapped to multiple IPs, while most hosts, 94.6% (635,709), are resolved to a single IP address. At the same time, the compiled dataset gives yet another support for the magnitude of virtual hosting: there are, on average, nine hosts per one IP address[4]. 77.2% (553,707) of all hosts in the dataset share their IPs with at least 20 other hosts.

After exclusion of 'redundant' host-IP pairs from the overall set (step 3 of Section 4.1), we left with 672,058 host-IP pairs (672,058 unique hosts on 78,736 unique IP addresses) in the main set S and with 45,182 host-IP pairs in the 'redundant' set S^*.

We then clustered 672,058 host-IP pairs by their IPs and, in a such manner, got 78,736 groups of pairs, each having from one to thousands of hosts. We formed three strata using the following stratification criteria: Stratum 1 (S1) included those host-IP pairs which IP addresses are each associated with seven or less hostnames, groups of size from 8 to 40 inclusive formed Stratum 2 (S2), and Stratum 3 (S3) combined groups with no less than 41 hosts in each. 8 and 41 were chosen to make S1 contain 90% of all IP addresses and to put 70% of all hosts into S3. Table 1 presents the numbers of IP addresses and hosts in each stratum. One can particularly observe that S3 comprises 70% (472,474) of all hosts and only 2% (1,860) of all IP addresses.

We randomly selected 964, 100 and 11 primary sampling units (IP addresses) from S1, S2 and S3 correspondingly. It resulted in 6,237 secondary units (hosts) in total to crawl (see also Table 1 for numbers across strata). Hosts of every sampled IP were crawled to depth three[5] as described in Section 4.2.

We calculated the estimates for the total numbers of deep web sites and databases and their corresponding confidence intervals according to the formulas given in Section 4.4. The final results are presented in Table 2. The '*Num of all*' column shows (in italic) the numbers of deep web sites and web databases that were actually detected in strata. However, not all of them were appeared to be Russian deep web sites. In particular, several sampled hosts in .RU were in fact redirects to non-Russian deep web resources. Another noticeable example in this category was `xxx.itep.ru`, which is one of the aliases for the Russian-mirror of

[4] A host resolved to multiple IPs is counted for each corresponding IP.
[5] Discussion on crawling depth value is given in [3].

Table 2. Approximate 95% confidence intervals for the total numbers of deep web sites (*dws*) and web databases (*dbs*) in each stratum and in the entire survey

Strata	**Num of all:**		**Num of Russian:**		**Num of Russian, corrected:**	
	dws	dbs	dws	dbs	dws	dbs
Stratum 1:						
- *Detected in sample*	*80*	*131*	*72*	*106*	*61.2*	*86.7*
- Conf. interval, [10^3]	6.0±1.4	9.9±3.6	5.4±1.3	8.0±2.8	4.6±1.2	6.6±2.1
Stratum 2:						
- *Detected in sample*	*38*	*46*	*38*	*46*	*36.1*	*44.1*
- Conf. interval, [10^3]	2.1± 0.7	2.5± 1.1	2.1± 0.7	2.5± 1.1	2.0± 0.7	2.4± 1.1
Stratum 3:						
- *Detected in sample*	*64*	*87*	*55*	*68*	*51.2*	*62.6*
- Conf. interval, [10^3]	9.6±3.4	13.0±5.5	8.2±3.5	10.2±3.5	7.6±3.6	9.3±3.9
Survey total, [10^3]	17.7±3.7	25.4±6.5	15.7±3.7	20.7±4.4	**14.2±3.8**	**18.3±4.4**

arXiv (`http://arxiv.org/`), an essentially international open e-print archive. We excluded all such non-Russian resources and put the updated numbers in the '*Num of Russian*' column. We also examined each deep web site on its accessibility via host(-s) on IP(-s) different from a sampled IP (a corresponding weight should be assigned to a deep web resource accessible via hosts on two or more IPs [15]) and aggregated the numbers in the '*Num of Russian, corrected*' column of Table 2.

The survey results, the **overall numbers of deep web sites and web databases in the Russian segment of the Web as of September 2006** estimated by the Host-IP clustering method are **14,200±3,800** and **18,300± 4,400** correspondingly.

5.1 Comparison: Host-IP Clustering Method vs. rsIP Method

To compare the Host-IP method with the rsIP we used the list of Russian deep web sites detected by the Host-IP technique, namely, 72, 38 and 55 deep web sites found in samples of S1, S2 and S3 correspondingly (see the '*Num of Russian dws*' column in Table 2). We compiled the list of IP addresses on which these sites are running (multiple IPs were added for those mapped to multiple IPs) and then applied the rsIP method to the list. The results are summarized in Figure 1, where the left chart shows how many deep web sites within each specified group of sampled hosts were detected by the Host-IP and rsIP methods, and the right chart depicts the overall estimates produced by both methods for the numbers of deep web sites in each stratum and in total. For instance, the rsIP and Host-IP applied to hosts (of sample S1) sharing their IP with one or two other hosts detected 10 and 15 deep web sites correspondingly. The outcome is quite expectable while deep web sites of S3 were mostly undetectable by the rsIP method, around two thirds of deep web resources in S1 and one third of resources in S2 were successfully recognized by the rsIP. The interesting

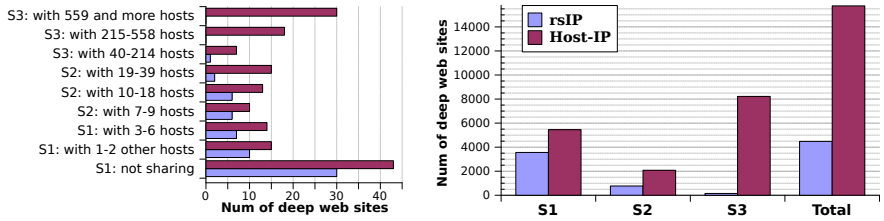

Fig. 1. Comparison of rsIP and Host-IP methods: *(left)* numbers of deep web sites detected in samples of strata (S1,S2,S3) among different types of hosts by rsIP and Host-IP; *(right)* numbers of deep web sites for each stratum and in total estimated by rsIP and Host-IP.

observation is that the rsIP is not efficient even for the hosts not sharing their IPs - 13 out of 43 deep web sites were overlooked by the rsIP[6] (see Figure 1(left)).

The factual estimates (approximate 95% confidence intervals) for the overall number of Russian deep web sites derived by the Host-IP and rsIP methods on the same IP list are 15,700±3,700 and 4,500±1,200 correspondingly. The rsIP approach therefore missed approximately seven out of ten deep web resources. In this way, **the rsIP** (used in previous deep Web characterization efforts) applied to our dataset **would be resulted in the estimates that are 3.5 times smaller than the actual figures**. The main impact to this difference is due to S3 with more than a half of all Russian deep web sites (8,200 out of 15,700), which are almost completely undetectable by the rsIP approach (see Figure 1(right)).

6 Related Work and Discussion

In Section 2, we mentioned existing deep Web surveys and discussed their limitations, where the most serious one is ignoring the virtual hosting factor. Several studies on the characterization of the indexable Web space of various national domains have been published (e.g., [5,11,18]). The review work [4] surveys several reports on national Web domains, discusses survey methodologies and presents a side-by-side comparison of their results. The idea of grouping hosts based on their IP addresses was used by Bharat et al. [8] to identify host aliases (or mirrored hosts according to Bharat's terminology). At the same time, we are unaware of any web survey study based on the Host-IP clustering approach.

One of the most surprising results in our survey is the fact that around a half of all deep web sites are hosted on IP addresses shared by more than 40 hosts (see 'S3' row of Table 2). It is somewhat unexpected since a deep web site serves dynamic content and thus normally requires more resources than an ordinary web site. Common sense suggests that a dedicated server (i.e., a server hosting from one or two to perhaps dozens of hosts, most of which are aliases) would be a better alternative for hosting a web site with database access. Nevertheless, it gives us just another strong justification for taking into consideration the virtual hosting factor.

[6] Empty results of reverse IP resolving were the reasons of overlooks.

7 Conclusions

We described a new sampling strategy, the Host-IP cluster sampling, that addresses drawbacks of previous deep Web surveys and accurately characterizes a large national web domain. We demonstrated the magnitude of virtual hosting and showed the consequences of ignoring it on a real dataset. We also compared our approach with the rsIP technique used in previous deep Web characterization studies and showed that the rsIP estimates for total number of deep web sites and databases are highly underestimated. Finally, we conducted the survey of Russian deep Web and estimated, as of September 2006, the overall number of deep web sites in the Russian segment of the Web as 14,200±3,800 and the overall number of web databases as 18,300±4,400.

References

1. April 2004 Web Server Survey (April 2004), http://news.netcraft.com/archives/2004/04/01/april_2004_web_server_surv%ey.html
2. DNS load balancing report (April 2004), http://www.securityspace.com/s_survey/data/man.200404/dnsmult.html
3. Baeza-Yates, R., Castillo, C.: Crawling the infinite Web: five levels are enough. In: Leonardi, S. (ed.) WAW 2004. LNCS, vol. 3243, pp. 156–167. Springer, Heidelberg (2004)
4. Baeza-Yates, R., Castillo, C., Efthimiadis, E.N.: Characterization of national Web domains. ACM Trans. Internet Technol. 7(2) (2007)
5. Baeza-Yates, R., Castillo, C., López, V.: Characteristics of the Web of Spain. Cybermetrics 9(1) (2005)
6. Bergman, M.: The deep Web: surfacing hidden value. Journal of Electronic Publishing 7(1) (2001)
7. Bharat, K., Broder, A.: A technique for measuring the relative size and overlap of public web search engines. Comput. Netw. ISDN Syst. 30(1-7), 379–388 (1998)
8. Bharat, K., Broder, A., Dean, J., Henzinger, M.: A comparison of techniques to find mirrored hosts on the WWW. J. Am. Soc. Inf. Sci. 51(12), 1114–1122 (2000)
9. Chang, K., He, B., Li, C., Patel, M., Zhang, Z.: Structured databases on the Web: observations and implications. SIGMOD Rec. 33(3), 61–70 (2004)
10. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: Proc. of WebDB 2004 (2004)
11. Gomes, D., Silva, M.J.: Characterizing a national community web. ACM Trans. Internet Technol. 5(3), 508–531 (2005)
12. O'Neill, E.T., McClain, P.D., Lavoie, B.F.: A methodology for sampling the World Wide Web. Annual Review of OCLC Research 1997 (1997)
13. Shestakov, D.: Deep Web: databases on the Web. In: Handbook of Research on Innovations in Database Technologies and Applications, pp. 581–588. IGI Global (2009)
14. Shestakov, D.: On building a search interface discovery system. In: Proceedings of VLDB Workshops 2009, pp. 114–125 (2009)
15. Shestakov, D.: Measuring the deep Web (2011) (submitted)
16. Shestakov, D., Salakoski, T.: On estimating the scale of national deep Web. In: Wagner, R., Revell, N., Pernul, G. (eds.) DEXA 2007. LNCS, vol. 4653, pp. 780–789. Springer, Heidelberg (2007)
17. Thompson, S.: Sampling. John Wiley & Sons, New York (1992)
18. Tolosa, G., Bordignon, F., Baeza-Yates, R., Castillo, C.: Characterization of the Argentinian Web. Cybermetrics 11(1) (2007)