THE SPANISH WEB IN NUMBERS
Main Features of the Spanish Hidden Web

Manuel Álvarez, Fidel Cacheda, Rafael López-García and Víctor M. Prieto
Department of Information and Communication Technologies, University of A Coruña,
Facultade de Informática – Campus de Elviña S/N, A Coruña, Spain

Keywords: Information retrieval, Hidden Web, Spanish Web.

Abstract: This article submits a study about the web sites of the ".es" domains which focuses on the level of use of the technologies that hinder the traversal of the Web to the crawling systems. The study is centred on HTML scripts and forms, since they are two well-known entry points to the "Hidden Web". For the case of scripts, it pays special attention to redirection and dynamic construction of URLs. The article concludes that a crawler should process those technologies in order to obtain most of the documents of the Web.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The "Hidden Web" or "Deep Web" (Bergman, 2000) is the portion of the Web that is not directly linked. Conventional crawlers cannot treat some of the technologies that constitute the entry points to those documents considered "hidden". On the one hand, forms are a way to access the server-side Hidden Web. On the other hand, technologies like scripting languages or Flash are entry points to the client-side Hidden Web.

Álvarez et al. (2009) started a high-level study about the ".es" domains, introducing quantitative statistics. This article continues that study by analyzing the content of the first page of each ".es" domain, in order to determine what Hidden Web technologies they use. The objective is to conclude if a crawling engine should deal with these technologies in order to obtain more documents.

The structure of this article is as follows. Section 2 reviews the related work. Section 3 explains the architecture of the crawler we have used. Section 4 analyses the results of our experiment. Finally, Section 5 concludes and explains the future works.

2 RELATED WORK

Most of the studies about the Web only deal with the Surface Web, but just a few also deal with the Hidden Web (Chang et al., 2004). There are also some Web sites that offer statistics about the indexed content (de Kunder, 2011), the number of servers (NetCraft, 2011) or the content of the web pages (BuiltWith, 2011) (Google, 2011). In addition, some organizations are in charge of both maintaining the domain names and counting the machines registered in them. (Internet Systems Consortium, 2011). Some of them make reports about the evolution of the domains (Verisign, 2011) (Red.es, 2011). However, there are not public reports that analyse the pages of the Spanish sites in order to determine the technologies they use.

There are also several works that reveal the difficulties that a crawler has to overcome in order to retrieve web documents. Scripting languages is an example, but Weideman and Schwenke (2006) and Wu and Davidson (2005) stated that crawlers often do not evaluate them although they are widely used.

3 ARCHITECTURE

Unlike conventional crawlers, the one used in this research does not follow the links of the web pages, but uses a list of domains to obtain their state and their main page. A Crawling Analysis Module obtains the data of the Álvarez et al. (2009) study. A Content Analysis Module has been added to compute and store the features of each web page in order to simplify the generation of content statistics. Figure 1 shows the system architecture:

Álvarez M., Cacheda F., López-García R. and M. Prieto V..
THE SPANISH WEB IN NUMBERS - Main Features of the Spanish Hidden Web.
DOI: 10.5220/0003626603630366
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 363-366
ISBN: 978-989-8425-79-9
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

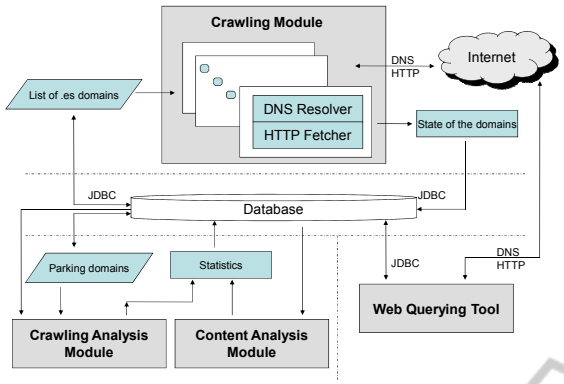

Figure 1: Crawler architecture.

Some elements like the "DNS resolver" were especially tuned to get extra information from DNS servers. On the other hand, the content analysis module uses the CyberNeko HTML analyser to treat every HTML document as an XHTML resource.

4 EXPERIMENTS

According to Álvarez et al. (2009), the Spanish Web had a total of 1,093,193 domains in May of 2009, but only 577,442 (52.82%) had a web server. From here on, we comment the results we obtained from the analysis of the main page of the ".es" domains with a web server in terms of including scripting languages (4.1), web forms (4.2) and other technologies (4.3).

4.1 Scripts

Scripts are the main entry point to the client-side Hidden Web. In the case of the Spanish Web, we have found scripts in 266,737 domains, 46.2% of those which had a server that did not return an error. We have also classified them in 542,322 invocations of external script files in 179,576 domains (31.1%), 744,111 <script> tags in 231,059 domains (40%) and 2,266,881 occurrences in HTML attributes inside 147,617 domains (25.6%).

Not all the scripts follow the recommendations of the RFC 4329 in order to make their language clear, but most of the scripts we have found follow the ECMAScript standard. In fact, almost all of them are written in JavaScript. VBScript have very few occurrences (1,769) and languages such as TCL script could not be found in the Spanish Web.

We have also studied which HTML tags contain the biggest number of <script> tags (including both embedded scripts and external calls). Scripts should be placed in the <head> tag or in the beginning of the <body> tag to ensure visibility, but Table 1 shows that some components that generate markup can contain <script> tags too.

Table 1: Location of the <script> tags inside other tags.

HTML tag	Number of scripts	Domains
<head>	530,522	200,713
	260,275	75,619
<body>	228,887	99,361
<td>	116,191	43,992
<p>	31,488	13,941

Table 2 shows the tags that contain the biggest number of scripts in their attributes.

Table 2: Location of <script> tags inside HTML attributes.

HTML tag	Number of scripts	Domains
<a>	1,305,831	95,402
<img>	196,419	18,880
<td>	184,657	6,495
	140,825	12,519
<input>	111,013	36,531

Most of the scripts are located in "onXXX" attributes of <a> links. In many cases, this is done to dynamically generate the target URL. However, we have found scripts in many other places such as the "action" attribute of the forms, etc. We have paid special attention to the "onLoad" event of the <body> tag, appearing in 47,064 domains (8.2%).

Regarding the number of scripts per page, Figure 2 shows the distribution:

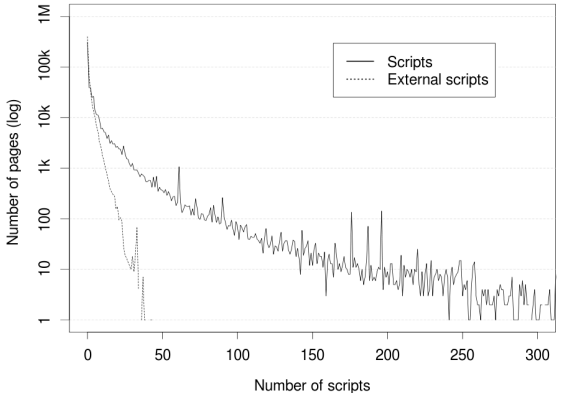

Figure 2: Use of scripts in the Spanish Web.

We have obtained a power law distribution for both the embedded scripts and the external invocations. This means that most of the pages do not include any scripts, a lot of them include very few and only some pages include a big number of them. We have also confirmed that they do not

follow a Zipf distribution. Figure 2 only shows the distributions of external scripts and all the scripts, since the distribution of embedded scripts is very similar to the latter. Only a small number of sites use more than 300 scripts, so we can omit a big portion of the tail of the distribution.

We have also found both sites that invoke the same scripts many times and sets of sites that invoke the same scripts. In the latter case, most of the times the pages of a set of sites have the same content (so they represent replicate content).

Furthermore, we have studied which script files are the most invoked. We have concluded that there is a group of 63 JavaScript files whose names appear more than 1000 times. Table 3 shows the top ten positions of that group with the number of occurrences and the number of domains:

Table 3: Script files that are invoked more times.

File name	Calls	Domains
show_ads.js	36,248	18,443
urchin.js	33,898	32,873
AC_RunActiveContent.js	29,746	28,526
swfobject.js	19,918	18,887
prototype.js	9,029	8,837
mootools.js	8,483	8,032
jquery.js	8,207	7,851
caption.js	5,947	5,916
scriptaculous.js	5,172	5,028
funciones.js	4,578	4,419

We have also tried to group those libraries by functionality and to count the number of domains that use libraries of each functionality. Table 4 shows the results:

Table 4: Functionality of the most common scripts.

Functionality	Domains	Calls
Management of Flash and active content	51,895	59,713
Visit count and generation of statistics	39,354	41,777
Content dynamization with AJAX	28,819	41,767
Content rendering and image treatment	22,185	24,598
Menu generation	4,714	5,198
Data treatment and validation	4,376	6,586

We have concluded that, although there are a lot of libraries on the Web, we can group them in a small number of functionalities (generation of statistics, AJAX, Flash, image treatment, etc.).

An interesting feature of the use of scripts is how they are used to create URLs dynamically. One of the cases that crawlers hardly ever manage well is when URLs have parameters whose values are injected with JavaScript, like in the following code:

```
location.href =
"http://www.tienda.es/prod?id="+id;
```

In many cases, crawlers will believe that the inner expression showed below is an URL:

```
"http://www.tienda.es/prod?id=" + id
```

Or if they perform a tokenization, it could be:

```
http://www.tienda.es/prod?id=
```

Actually, the result of the tokenization is an URL, but as the product identifier is missing, it probably will not lead the crawler to the expected resource. The problem here is that we cannot guess the best values for the parameters unless we interpret JavaScript, but other techniques could be researched. Table 5 shows the number of URLs we found in typical redirection sentences, as well as the number of simple "potential" URLs that have parameters injected by means of JavaScript as in the previous example. Moreover, it also shows how many of them are considered "well formed" by the algorithm of the OpenSource crawler Nutch, which uses, first, regular expressions to detect potential URLs and, then, a filter to discard not valid ones.

Table 5: Finding complete and potential URLs.

Name	Number	Pass Nutch filter	%
Complete URLs	41,716	41,590	99.7
Potential URLs	8,709	8,581	98.5

As it is shown in Table 5, we have found 8,709 potential URLs that could be completed with some extra processing. However, conventional crawlers would treat them as valid URLs although they actually point to error pages or uninteresting pages.

4.2 Web Forms

Forms are the main entry point to the server-side Hidden Web, so we need to study them thoroughly. We have found 188,712 forms in 124,865 domains (21.6%) following a power law distribution. 122,417 of them (64.9%) make their request by POST and 48,443 (25.7%) use GET for that purpose. Also, 17,779 forms (14.2%) do not specify a method, so the default value for them is GET too.

Table 6 shows the use of password fields in forms. The percentages are relative to the number of pages with forms. These fields are often associated to authentication, register or password change tasks.

Table 6: Use of password fields.

Password fields	Forms	Domains	%
1 (authentication)	26,918	25,832	20.7%
2 (register)	251	239	0.2%
3 (password change)	33	33	<0.1%

We have also found 5,346 forms with only one text field in 4,861 domains (3.9%). They are typically used for doing simple searches. Regarding more complex searches, 34,006 forms in 31,288 domains (25%) are made of several text fields and buttons, and 126,095 forms in 91,235 domains (73%) contain at least two elements of the following categories: <input>, <select> and <textarea>.

4.3 Other Technologies

Some tags like <meta> can contain interesting information for crawlers. Table 7 shows some of the purposes for which they were used:

Table 7: Use of <meta> tags.

Functionality	Domains	%
Refresh or redirection	22,064	3.8%
Refresh (no redirection URL)	1,346	0.2%
Robot Excusion Standard	128,288	22.2%
Keywords	246,752	42.8%
Sending cookies	26	<0.1%

Although stating the keywords and Robot Exclusion Standard (Koster, 1994) are the most popular purposes, many search engines do not take the first one into account since Gyöngyi and Garcia-Molina (2005) explained how they could be used to do *boosting* (an unfair rise of the score of the page).

Flash applications are another difficulty that crawlers have to overcome. We have found <object> tags in 89,911 domains (15.6%). 25.060 of them (4.3%) did not have <a> links, so most of them were 100% Flash sites.

5 CONCLUSIONS

This article has shown the main results of an analysis of the ".es" websites in 2009. Particularly, 15.6% of the domains had <object> tags, 21.6% had forms and 46.2% contained scripts, most of them written in JavaScript. Other scripting languages are rarely used, so only the effort of processing JavaScript would be justified in order to retrieve more pages from the Web.

The future work would consist in making new crawlings of the Spanish Web in order to supervise the evolution of the documents and the technologies.

ACKNOWLEDGEMENTS

This research work has been financed by Ministerio de Educación y Ciencia of Spain and the FEDER funds of the European Union (Project TIN2009-14203) and the list of ".es" domains that was employed in the crawling is courtesy of Red.es.

REFERENCES

Álvarez, M., Cacheda, F., Pan, A., 2009. Análisis Macroscópico de los Dominios .es. In *JITEL'09, VIII Jornadas de Ingeniería Telemática.*

Bergman, M., 2000. The Deep Web. Surfacing Hidden Value. In *Technical Report*, BrightPlanet LLC.

BuiltWith Trends. 2011. BuiltWith Technology Usage Statistics. In *http://trends.builtwith.com/*

Chang, K. C.-C., He, B., Patel, M., Li, C., Zhang, Z., 2004. Structured Databases on the Web: Observations and Implications. In *SIGMOD Record, vol. 33, no. 3.*

De Kunder, M. 2011. The size of the World Wide Web. In *http://www.worldwidewebsize.com/*

Google. 2011. Web Authoring Statistics. In *http://code.google.com/intl/es-MX/webstats/index.html*

Gyöngyi, Z., Garcia-Molina, H., 2005. Web Spam Taxonomy. In *AIRWeb'05, Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*.

Internet Systems Consortium. 2011. The ISC Domain Survey. In *http://www.isc.org/solutions/survey*

Koster, M. 1994. A Standard for Robot Exclusion. In *http://www.robotstxt.org/orig.html*

Netcraft. 2011. March 2011 Web Servers Survey. In *http://news.netcraft.com/archives/category/web-server-survey/*

Red.es. 2011. In *http://www.red.es*

VeriSign. 2011. Internet Profiling Service Statistics. In *http://www.nic.at/en/uebernic/statistics/ips_statistics_informations/*

Weideman, M., Schwenke, F. 2006. The influence that JavaScript™ has on the visibility of a Website to search engines - a pilot study. In *Information Research, vol. 11, no. 4*.

Wu, B., Davison., B.D. 2005. Cloaking and Redirection: A Preliminary Study. In *AIRWeb '05, Proceedings of First International Workshop on Adversarial Information Retrieval on the Web*.