The Evolution of the Web

Víctor M. Prieto*, Manuel Álvarez* and Fidel Cacheda*
*Department of Information and Communication Technologies, University of A Coruña, Coruña, Spain 15071
Email: {victor.prieto, manuel.alvarez, fidel.cacheda}@udc.es

***Abstract*—In recent years several studies have been carried out which have attempted to characterise the Web on different levels, and within a specific period of time. This paper presents a study on the evolution of the Global and Spanish Web in the last 3 years. It will analyse the main features of the Web, grouped on different levels and during three separate periods of time. The aim is to obtain information on changing trends on the Global and Spanish Web over the years, comparing and analysing the results of these trends. This has been conducted paying special attention to such aspects as the degree of similarity of the Web, the evolution of the age of web pages and the most commonly used web technologies.**

I. Introduction

The WWW can be considered as the largest repository of documents ever built. According to a study presented by Gulli and Signorini [1] in 2005, the Web consists of billions of web pages. Due to its large size, search engines are essential tools for users who want to access relevant information relating to a specific query. Search engines are complex systems that allow, among other things: gathering, storing, managing and granting access to the information. Crawler systems are those which perform the task of gathering information. These programs are capable of traversing and analysing the Web in a certain order, by following the links between different pages.

The task of a crawling system presents numerous challenges due to the quantity, variability and quality of the information that it needs to collect. Among these challenges, specific aspects can be highlighted, such as the technologies used in web pages to access to data, both in the server-side [2] or in the client-side [3]; or problems associated with web content such as Web Spam [4] or repeated contents [5], etc. To get a detailed enumeration it is necessary to analyse the Web.

This study presents an analysis of the most important features of the Web and its components and its evolution over a period of time. Particular emphasis is placed on aspects such as the similarity and age of the pages or the use of client/server side technologies. The study focuses on a snapshot of the Spanish and Global Web for three different years: 2009, 2010 and 2011. The results for each year are analysed independently and together to facilite the study of both the features at any given time and the changes between the different analysed years. The objective of the study is to characterise the Web and to determine how its changes evolve.

The structure of this paper is as follows. In Section II we comment works related with the study of the Web. Section III shows the methodology used along the study. Section IV explains the dataset used in this paper. In Section V we analyse the obtained results on the Global and Spanish Web, and discuss them to show the evolution of the Web. Finally, Section VI includes our conclusions and the possible future work in this field.

II. Related Work

The characterisation of the Web is a topic widely studied in the supported literature. Baeza-Yates *et al.* in [6] performed a study which analyses various features of the Web at several levels: web page, web site and national domains. On the other hand, there are several studies that are focused on the Web of a particular country. In 2000, Sanguanpong *et al.* [7] presented an analysis of various issues related to web servers and web documents in Thailand. Baeza-Yates *et al.* presented two papers [8] and [9], which were focused more specifically on the characteristics of the Spanish and Chilean Web, respectively. The Spanish Web was also studied by Prieto *et al.* in [10]. In 2002, Boldi *et al.* [11] presented an interesting paper, where the authors studied differents features (content and structure analysis, web graph, etc.) of the African Web. Gomes *et al.*, carried out a study to characterise the community Web of the people of Portugal [12]. The authors studied differents features such as: the number and domain distribution of sites, the number and size distribution of text documents, the structure of this Web, etc. Years later, Miranda and Gomes [13] performed a study which presents trends on the evolution of the Portuguese Web, derived from the comparison of two characterisations of a web portion performed within a 5 year interval. This study analyses several metrics regarding content and site characteristics. Modesto *et al.* [14] presented an paper, which analyses the features of approximately 2% of the .br domains. The results have been compared with the results obtained in other studies on the Chilean and Greek Web. Finally, another similar study was performed by Efthimiadis and Castillo [15], where the authors did a characterisation of the Greek Web.

On the other hand, there are studies that focus on studying a specific feature of the Web. It is the case of the study presented by Grefenstette and Nioche [16], where the authors analysed the English and non-English language used on the Web. A relevant study was the performed by Bharat *et al.* [17], which discussed the links between Web sites and its meaning. Another study focused on a particular feature of the Web, was the performed by Downey [18], where the author analysed models for web page sizes. There are other studies that focus exclusively on the structure of the Web, such as that conducted by Broder *et al.* [19]. In 1999, Huberman and Adamic [20] carried out a study where the authors characterise the distribution of web pages per web site. Other specific feature is the Deep or Hidden Web (pages that are accessed through web forms or by means of client-side technologies). Among the studies focused on it, we can highlight [3] and [21].

There are several studies with respect to the dynamic and age of Web pages. The most notably of these is that presented by Lewandowski [22], which discusses the evolution of the age of the pages over several years. Fetterly *et al.* [23] included a study about the degree of change of each page, and which factors are correlated with change intensity.

Numerous studies have examined the Web from different points of view. However, none of them have studied the evolution of the main features of the Spanish and Global Web through time. Moreover, this paper studies several interesting features along the time, such as similarity and age of the pages or the use of client/server side technologies, not analysed in other studies of the Global or national Web.

III. Methodology

The analysis of the web can be performed at various levels of granularity [24].Below we describe the levels of web analysis included in this work, together with the characteristics analysed in each of them.

- Word: The study of this level will provide data about the vocabulary used on the Web, the stopwords and the most commonly used HTML tags.
- Web content: The analysis of this level will allow us to obtain data about the evolution of the content size and its relationship to the useful content. This will also increase our knowledge about the evolution of the most widely used languages on the Web. Finally, the analysis of this level will provide data about how the formats of multimedia files evolve on the Web.
- Web page: At this level we analyse the characteristics of an entire web page, the length of the URLs and the level of compression of the pages. In addition, we will pay special attention on the age and the similarity of web pages.
- Web site: At this level we will discuss the main features of web sites, defined as collections of related web pages common to a domain or subdomain. We will analyse the number of links (inlink, outlink, static, dynamic, relative and absolute) on the Web.
- National Web: The analysis of this level will allow us to categorise the software used by web servers and the evolution in the amount of new/removed domains on the Spanish Web.
- Global Web: As for the national Web, we will show an analysis of the use and evolution of the software used by web servers.

Across all levels studied, we will discuss the results obtained for the Global and Spanish Web, and we will compare them to find their differences and similarities.

IV. Datasets used for Testing

In this section we describe the two datasets used for the study, and how we obtained each one.

We want to clarify that this study analyses the Global and Spanish Web. So, we wish to know the similarities and differences between the Global and Spanish Web to adapt the features and policies of a Global web crawler to characteristics of Spanish Web. We think that the data, crawling policies and conclusions obtained in this paper are very useful to other researchers that want to study other different national Web or to develop a crawler for a specific national Web.

To generate the Spanish dataset we have selected the web sites under the .es domain, regardless of the language used in their sites or the country where IP of the web server was located. We obtained the complete list of the .es domains due to collaboration agreement with the Spanish Business Public Entity called Red.es[1] (Red.es is responsible of managing all the .es domains). For our dateset we have not focused on web sites of a specific topic, type (academic, business, etc.) or language (the study does not deal exclusively with pages written in Spanish). On the other hand, the dataset does not contain web pages of the social media Web (eg, Vimeo or Flickr), which is constantly growing, but from our point of view this type of Web does not represent the Web of a country. This is because platforms like Vimeo or Flickr use the same formats, styles and technologies, independently of the country of the user.

With this system we carried out 3 crawls from 2009 until 2011, one per year, to obtain 3 datasets of the Spanish Web. The crawling system was configured to not process sites external to the .es domain. This process was repeated from 2009 until 2011, compiling 3 datasets: a) 2009 with 577,000 documents, b) 2010 with 785,000 documents and c) 2011 with 1,050,000 documents. Upon completion, we had obtained a full dataset of 2,412,000 web pages.

For the purpose of analysing the Global Web, we have used the data provided by "The Stanford WebBase Project"[2], which is part of "Stanford Digital Libraries Project"[3]. The full dataset contains more than 260 TB of data, organized into subsets of different topics (general thematic, natural disasters, governments, etc.). Our dataset for the Global Web only considers the general thematic subdataset. Of the more than 700 million obtained pages, we chose a random sample to get a subset of 10 million pages for each year, 2009, 2010 and 2011. Overall, the global dataset contains approximately 30 million web pages.

Finally, we want to remark that these two datasets are comparable, although they have been created using different crawlers, one implemented by us and another at Stanford, due to the purpose and the strategy used by both crawlers is the same. Moreover, the types of web sites and web pages (business, academic, blogs, etc.) contained in both datasets are similar and contain web pages written in any language.

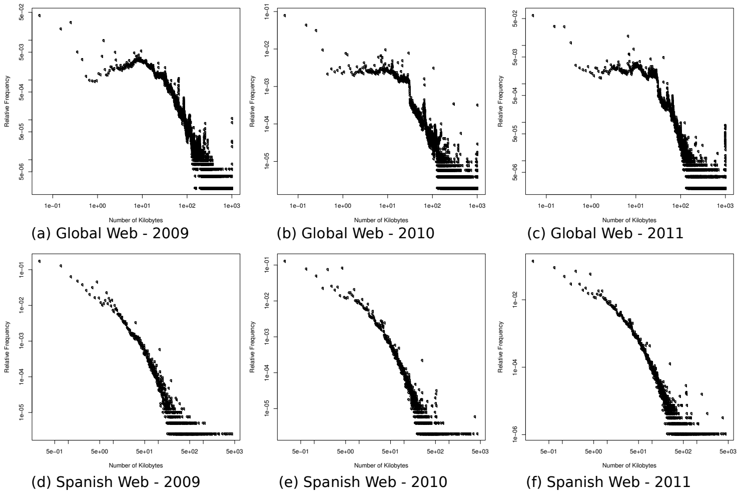
V. Results of the Evolution of the Web and its Implications

Taking the defined datasets as a starting point, in the following sections the results obtained for each mentioned characteristics will be discussed.

[1] http://www.red.es

[2] http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/

[3] http://diglib.stanford.edu:8091/



(a) Global Web - 2009 (b) Global Web - 2010 (c) Global Web - 2011

(d) Spanish Web - 2009 (e) Spanish Web - 2010 (f) Spanish Web - 2011

Fig. 1: Content size of a web page

A. *Features at Word Level*

Within the scope of web pages, a word may be used mainly as a term or as an HTML tag. This section analyses the words from these two points of view.

1) Vocabulary: In order to study the vocabulary used on the Web, we considered a word as any alphanumeric string with length larger or equals than one character. For both the results of the Global Web and the Spanish Web, we found that the number of pages containing terms with small Tf-Idf is increasing. This indicates a growing number of common words in web pages. So, selecting relevant web documents for a given set of terms, will be more difficult.

2) HTML tags: Another type of "important" words on the Web are HTML tags which create and shape web pages. Analysing the Global and Spanish Web, we observed that the 50 most used tags are common over the 3 years analysed (except for minimal changes). This shows us that has been little change in the way of developing web pages. On the other hand, we found many incorrect tags that hinder the task, causing the browser to display the page incorrectly, and perhaps interfering with the correct processing being carried out by the crawling system. In our opinion, these results mean that design of the web sites and web page have not changed in the last 3 years.

B. *Features at Web Content Level*

This section discusses the evolution of the total/useful size of the web pages, the most commonly used charsets, languages, music, video and picture formats, and an analysis of certain attributes of the "meta" tag.

1) Size of the total/useful content: The useful content of a web page is the main content, where the information is really placed, without HTML tags, links, images, etc. The useful content is used by search engines to provide the correct web documents to the user. An important fact for search engines and crawlers is the size of the downloaded and stored content, and its relation to the useful content of each page.

The process of extraction of the useful content of the web pages is very complicated. In our experiments, we have followed the approach developed by Donghua *et al.* in [25]. This study is based on that the location of the main content is very centralized and has a good hierarchical structure. The authors found that the threshold values of the DOM nodes [26] with useful content are obviously different from that of other DOM nodes in the same level. With these values, they have proposed an algorithm that judges the content by several parameters in the nodes (Link Text Density, Link Amount, Link Amount Density and Node Text Length).

In Figure 1 we can see the obtained results. It is important that, for this study, we have considered the full content of the pages (unlike other existing studies that truncate the pages to a certain size [9]).

In the Global Web (Figure 1 a, b, c), in the year 2009 the average content per page was 28.18 KB. This number has decreased in 2010 to 24.1 KB and in 2011 to 21.4 KB. This is a decrease of 24.06% in 3 years. Analysing the results taken for useful content, it is notable that in 2009 the average size of useful content was 6.49 KB, a number that does not change much in subsequent years. This indicates that the size of useful content in web pages has not changed over the years and that the relationship between total and useful content represents, in 2011, approximately 28% of the total.

Web pages of the Spanish Web (Figure 1 d, e, f) had in 2009, an average size of 9.98 KB, in 2010 and 2011 this figure grew to 11.8 and 13.4 KB, respectively. We observe that in these 3 years most of the pages were between 10 and 500 KB in size, and that there are some cases of very large sites that can reach up to 5 MB. We have also studied the useful content of the pages and their relation to the total. As regards the useful content, in 2009 we see that the average size of the useful content was 5.52 KB. In 2010 this figure grew to 6.27 KB and to 6.31 KB in 2011. In 2011 the useful content versus the total represented 47.08%.

This indicates that the average size of total content of Spanish web pages is less than half the content of other pages on the web. We have observed that the average size of total content on the Global Web is decreasing, and on Spanish Web it is increasing. This could be, as Downey discusses in its paper [18], due to the differences in the usage of HTML coding for writing Web pages and because the web pages tend to be more complex.

Comparing the obtained results with the published in the supported literature, we can see that the web pages of the Spanish and Global Web have more useful content than the web pages of Portuguese Web [12]. Concretely, in 2009 the Spanish and Global Web had 5.52 KB and 6.49 KB of useful content, respectively, and the Portuguese Web 2.8 KB [12], that is, approximately the half.

In short, we have observed that, although the size of web pages on the Global Web is decreasing and on the Spanish Web is growing, in 2011 the average size on the Global Web was approximately the double. Taking into account the useful content, we see no difference between the average size of the contents of studied Webs. However, the relationship between useful content and total content is smaller on the Global Web than on the Spanish Web, which represents about 50% of the total. This fact indicates that the web pages on the Global Web include a lot of HTML code. It is likely that these web pages use client-side technologies such as JavaScript to improve the user experience.

2) Language: In order to identify the language used in each web page, we have used the "language detector" library [27], which is based on Bayesian filters. It has a precision to detect the 53 supported languages of 0.99.

The results obtained for the Global and Spanish Web, are shown in Table I. On the Global Web, it is observed that the predominant language is English with 96.86% in 2009, decreasing to approximately 94% in 2010 and 2011. In the study presented by Grefenstette and Nioche [16] in the year 2000, the authors have estimated that about 70% of the web pages are written in English. According with our study this figure has increased 16.86%. In the last years the use of English language has decreased a bit, this is due to the increment of use of Spanish in web pages. Concretely, Spanish has increased from a 1.65% in 2009 to 2.54% in 2011.

On the Spanish Web, we note that in 2009 the Spanish language represented 63.08% of the total, and English 28.35%. However, curiously, we can see that from 2009 to 2011 the use of Spanish language has decreased by about 3% versus English. This may be due to the opening of new markets by the Spanish economy, which has led to an increasing number of contents being written in English. The other languages present in the Spanish Web in 2011 represent 8.57%, among which are the main European languages (Italian, German, French, Portuguese, etc.).

In 2011, the ratio of web pages written in Spanish is around 63%. However, the fraction of pages written in Portuguese on the Portuguese Web is around 73% [12], 75% of Brazilian pages in Portuguese [28] and with 90% written in Spanish on the Chilean Web. Analysing the results of other national domains in which English is the most used language on-line, such as the Web of Thailand [7] with 66% of web pages written in English, and several African [29] countries with 75%. Based on these percentages, we can see that the Spanish Web presents more languages than the national Web of other countries.

Web	Year						
		Language					
		en	**es**	**others**			
Global	**2009**	96.86%	1.65%	1.49%			
	2010	94.89%	2.36%	2.75%			
	2011	94.45%	2.54%	3.01%			
Spanish	**2009**	24.75%	66.37%	8.88%			
	2010	27.72%	62.59%	9.69%			
	2011	28.35%	63.08%	8.57%			
		Image file formats					
		Gif	**Jpg/Jpeg**	**Png**	**Ico**		
Global	**2009**	86.01%	9.83%	4.00%	0.06%		
	2010	90.94%	14.18%	4.80%	0.08%		
	2011	74.95%	14.36%	10.61%	0.09%		
Spanish	**2009**	62.23%	29.06%	7.65%	0.06%		
	2010	39.23%	35.77%	24.97%	0.02%		
	2011	45.42%	34.50%	20.02%	0.06%		
		Video file formats					
		Wmv	**Mov**	**Avi**	**Mpeg**	**Qt**	
Global	**2009**	59.50%	37.36%	2.74%	0.32%	0.08%	
	2010	71.46%	26.22%	1.90%	0.34%	0.07%	
	2011	54.10%	42.54%	2.23%	1.05%	0.08%	
Spanish	**2009**	76.56%	12.67%	10.37%	0.40%	0.00%	
	2010	62.92%	23.25%	13.47%	0.37%	0.00%	
	2011	70.27%	17.31%	11.00%	1.37%	0.07%	
		Format styles					
		Bold	**Italics**	**Underline**	**H1**	**H2**	**H3**
Global	**2009**	60.96%	19.46%	2.39%	3.64%	6.20%	7.36%
	2010	61.86%	17.32%	2.10%	5.25%	7.21%	6.26%
	2011	54.97%	16.83%	2.34%	6.40%	11.17%	8.28%
Spanish	**2009**	60.79%	5.89%	2.84%	8.41%	14.55%	7.52%
	2010	54.92%	7.17%	1.18%	9.51%	15.33%	11.89%
	2011	56.51%	7.16%	1.34%	8.62%	14.72%	11.65%
		Used charsets					
		ISO-8859-1	**UTF-8**	**ISO-8859-15**	**Windows1252**	**Others**	
Global	**2009**	57.54%	34.68%	0.57%	5.51%	1.91%	
	2010	43.23%	43.75%	0.64%	7.52%	4.86%	
	2011	38.71%	51.51%	0.61%	7.15%	2.02%	
Spanish	**2009**	61.50%	27.73%	0.53%	6.92%	3.32%	
	2010	31.82%	63.34%	0.60%	3.56%	0.68%	
	2011	41.07%	51.53%	1.25%	4.42%	1.73%	
		Other document types					
		Pdf	**Doc**	**Xml**	**Ppt**	**Ps**	**Txt**
Global	**2009**	80.21%	4.60%	6.88%	0.69%	0.10%	7.52%
	2010	80.04%	7.31%	2.19%	0.94%	0.09%	9.43%
	2011	82.87%	7.91%	5.16%	0.99%	0.06%	3.03%
Spanish	**2009**	84.47%	4.37%	9.30%	0.46%	0.00%	1.39%
	2010	84.67%	3.25%	10.73%	0.35%	0.01%	0.99%
	2011	96.86%	3.23%	8.64%	0.30%	0.01%	0.96%
		Server-Side technologies					
		Php	**Asp**	**Jsp/Jhtml**	**Perl**	**Shtml**	**Cgi**
Global	**2009**	25.17%	15.66%	22.00%	3.33%	17.66%	16.18%
	2010	21.62%	55.35%	5.86%	1.32%	13.72%	2.12%
	2011	31.99%	51.04%	6.54%	1.12%	8.09%	1.22%
Spanish	**2009**	73.08%	21.26%	2.74%	1.13%	1.01%	0.77%
	2010	76.11%	19.17%	1.09%	0.96%	2.14%	0.52%
	2011	71.65%	23.34%	1.72%	1.07%	1.79%	0.44%
		Client-Side technologies					
		JavaScript	**Flash**	**VbScript**	**Others**		
Global	**2009**	93.63%	6.28%	0.08%	0.01%		
	2010	95.83%	4.12%	0.05%	0.00%		
	2011	97.15%	2.84%	0.01%	0.00%		
Spanish	**2009**	70.67%	28.91%	0.41%	0.01%		
	2010	81.06%	18.81%	0.11%	0.02%		
	2011	77.01%	22.77%	0.20%	0.02%		

TABLE I: Results of: languages, image/video file formats, format styles, documents types and client/server-side technologies

3) Music file formats: We have studied the music file formats used on the Web. We have observed that during the 3 years studied the distribution of the music file formats has not changed notably. Focusing on the data obtained in 2011, both for the Global and Spanish Web, the most widely used format is MP3 with 92% and a 86.06%, respectively. The next most commonly used music format is WAV with a 4.39% and a 6.4%, on the Global and Spanish Web, respectively. Other music formats are WMA with 5.71% on the Spanish Web, and MIDI or ASF, with very insignificant percentages. The limited evolution that has been observed is because the MP3 format has good quality with a relatively small size, and because no new formats have appeared in the recent years that have been able to replace the MP3 format.

4) Image file formats: Table I shows the changes in the use of image file formats.

In the Global Web, there is a predominance of the GIF format, although it has been slowly decreasing through the years, from a 86.1% in 2009 to 74.95% in 2011. After GIF, the most used image formats are JPG and PNG, used in 2011 with a incidence of 14.36% and a 10.7%, respectively. Spanish Web uses more image file formats than the Global Web. Despite this, the GIF format is still the most common with a 45.42%, although JPG and PNG have been gaining ground. In 2009, 29.06% of pages used JPG versus 34.5% in 2011. The same thing happened to PNG, whose usage increasing from 7.65% in 2009 to 20.02% in 2011.

5) Video file formats: Table I shows the most used video formats. In the Global Web, we note that in 2009 the most commonly used format was WMV with 59.5%, followed by MOV with 37.36% and AVI with 2.74%. In 2010, WMV increased its presence until 71.46%. In 2011, the difference between WMV and MOV decreased, with 54% and 42.54%, respectively.

On the Spanish Web, we also see that the predominant format is WMV with a 76.56% in 2009 and a 70.27% in 2011. This percentage lost is due to the increment in the use of MOV, which in 2011 was used in a 17.31% of cases. The use of AVI format, does not vary greatly in the 3 studied years, remaining above 10%. The AVI format was one of the oldest video formats. It has good quality but it is very heavy. In previous years it was the most widely used format on the Internet. Today, WMV has similar quality but with a smaller size. Due to this, AVI has almost disappeared with other smaller formats taking its place. In summary, we see that on the Global Web the use of the different video formats is more distributed than the Spanish Web, which mainly uses WMV.

6) Styles: Search engines also consider the style in which certain terms of web page content are written. The fact that a word is highlighted (bold or italic) can indicate that it is more relevant than others. We have conducted a brief survey of commonly used standard styles. The results are shown in Table I. On the Global Web, we see that the most widely used style is bold, with a 60.96% in 2009, but that has decreased about 5% in 2011. The use of italics has also dropped approximate a 3% since 2009, decreasing from 19.46% to 16.83%. Perhaps, the decrease in bold and italics is due to the increment of H1, H2 and H3 tags, from 2009 to 2011. Analysing the results of the Spanish Web, we see that the most commonly used style is bold with 56.51%, followed by the styles of the title sections, H1, H2 and H3, with a 8.62%, 14.72% and 11.65%, respectively. These results seem logical because a text can have many words in bold, but less in H3, H2 and H1 styles.

In summary, we observed that on both the Global and Spanish Web, the most common web style is bold. We also see that the italics are used more than double on the Global Web than on Spanish Web, but the Spanish web uses more the styles referring to title sections (H1, H2 and H3).

7) Meta tags: An important part of the information on the web page content is in the attributes of the HTML "meta" tags, which are placed at the beginning of the HTML code and provide information to the user, browser, crawlers and search engines. There are several attributes of the "meta" tag, however we have only considered the two most relevant to perform the analysis.

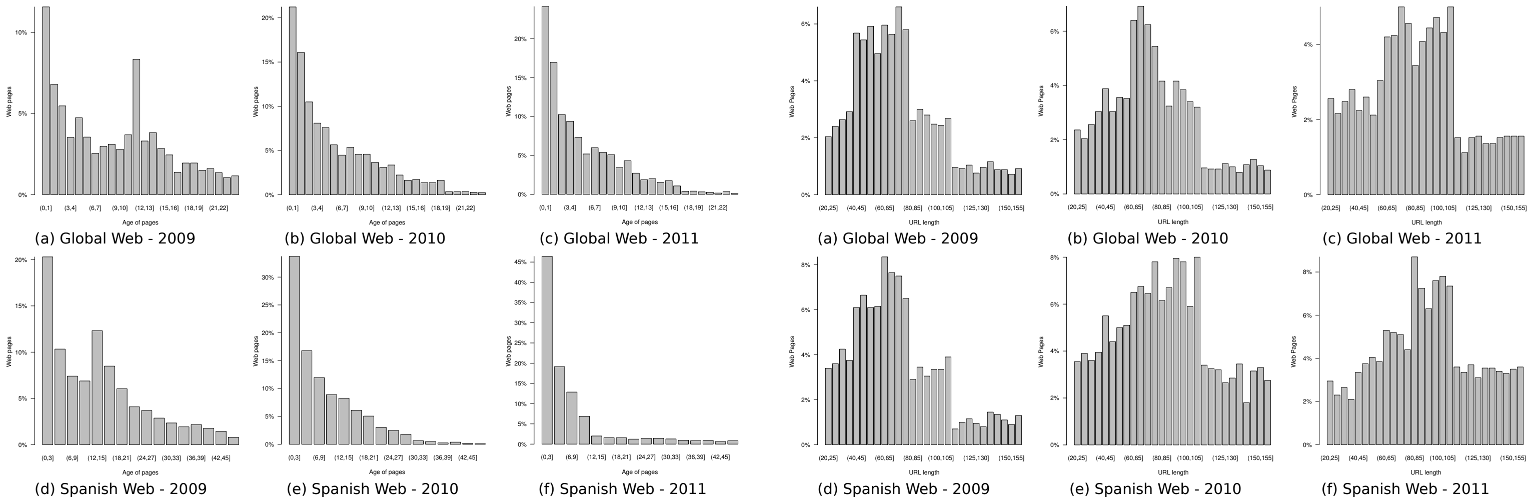
- Refresh: This attribute indicates the time when the content of the page should be updated. On the Global Web (a) in 2009 the 1.6% of the pages used the attribute "refresh". In 2010 and 2011 this result decreased to approximate 1%. On the Spanish Web, we found that the use of the "refresh" attribute is more common. In 2009 and 2010 the percentage of use was similar, about 4%. In 2011 the value rose to 4.9%.
- Content-type: Attribute that indicates the content type and character set used for coding the web page. The obtained results are shown in Table I. On both the Global and Spanish Web, we observe that in 2009 the most used charset was ISO-8859-1. In 2009, the use of ISO-8859-1 represented 57.74% on the Global Web and 61.5% in Spanish. In the years 2010 and 2011 the use of ISO-8859-1 fell to the 38.71% on the Global Web and 41.07% on the Spanish Web. This decrease occurred by the increment of UTF-8, which in 2009, reached 34.68% on Global Web, and 27.73% on Spanish Web. However, in 2010 and 2011 it has increased its presence to 51.51% on the Global Web and 51.53% on the Spanish Web. The increment in the use of UTF-8 and the decrement of ISO-8859-1, is due to the need for new types of coding that allow multilingual support.
- Keywords The keywords attribute of the "meta" tag, includes words which describe the content of the page. We have analysed the average number of keywords on the Global and on the Spanish Web. In the first case, we have observed that in 2009 the average was 8.37 keywords per web page, in 2010, 9.21 and in 2011, 8.89 keywords. We see that over the 3 years the number of keywords has remained at around 9. In the case of the Spanish Web, the average number of keywords has changed little over the years studied, remaining at 15 keywords per page. Comparing both results we see that the Spanish web pages use more keywords than the pages of the Global Web. This may be caused by the language itself, because certain languages tend to use more words than others.

C. Features at Web Page Level

In this section we will focus on analysing the characteristics of a web page. We will pay special attention to the length of the URL, the age of the pages and their similarity.

1) Other document types: One point that characterises web pages is the different types of documents that they contain. Table I shows the results for the most relevant file types.

Analysing the results for the Global and the Spanish Web, the type of document which appears more often is PDF. In 2011, around the globe 80.87% of all documents were PDF and on the Spanish Web that figure was 86.86%, i.e., 5% higher than on the Global Web. On the Global Web, the next most common type of document is the "Microsoft Word Document", which grew from 4.6% in 2009 to 7.91% in 2011. The textual file format was used 7.52% in 2009, and its usage has dropped to 3.3% in 2011. Finally, the use of XML documents in 2010 decreased, from 6.98% in 2009, to 2.91% in 2010, but increased again in 2011 to 5.16%. On the Spanish



(a) Global Web - 2009

(b) Global Web - 2010

Fig. 2: Evolution of the age and the length of URLs on Global and Spanish Web

Web, XML and DOC documents appear less than PDF, with an incidence of 8.64% and 3.23% in 2011, respectively. The fact that Microsoft Windows is the most used operating system, helps to increase the amount of Office documents, such as the DOC files.

In other countries PDF also is the document format most used in the Web. Concretely, on the Brazilian Web [14] it appears a 48%, approximately a 63% on the Chilean Web [9] and, 46% on the Web of Portugal [12]. In short, the obtained results are logical, since PDF documents, as its initials indicate (Portable Document Format), can be used in any operating system.

2) URL length: Figure **??** (a, b, c) shows the length in bytes of the URLs on the Global Web. In 2009, the average length was 70.88 bytes. In the years 2010 and 2011, the average length has increased to 77.2 and 84.66 bytes, respectively. Analysing the evolution, we note that the length of URLs has grown and also that the number of URLs with a length between 100 and 150 bytes is increasing.

The results for the Spanish Web are shown in Figure **??** (d, e, f). In 2009, many of the web pages had URLs between 45 and 75 bytes in size, and a small part of the Web had URLs with more than 100 bytes. In 2010, most of the pages have URLs between 65 and 100 bytes in size. Finally, in 2011 the figures show that most of the pages have URLs between 80 and 110 bytes and the group of web pages with URLs between 100 and 150 bytes has increased. On both the Global and Spanish Web, the length of URLs has grown since 2009. This is due to the growth of dynamic pages, new technologies and the need in many cases to send parameters within the URL. This change in the length of the URLs should lead to changes in the design of storage and caching URL systems for the crawlers, as well as in the queues to visit and for visited URLs.

Analysing the results presented in others studies, we have observed that the URL length on Spanish Web is similar to the observed on Portuguese Web, which has an URL length average of 62 characters. On the other hand, comparing our results on the Global Web with the presented by Suel and Yuan [30], we have observed that the URL length average was lower in 2001, 50 characters, than in 2011.

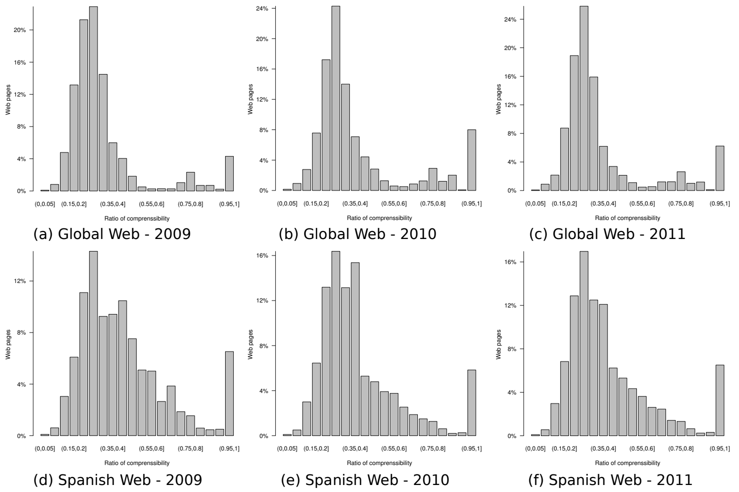
3) Age: It represents the time validity of a web page. We have used the “Last-Modified” HTTP header to know when a web page has been modified and therefore to know its age. Due to the “Last-Modified” header of the web pages dynamically generated is always set when a page is generated, we have only selected those web pages of the dataset that are not dynamic.

Figure **??** (a, b, c) shows the results for the Global Web. In 2009 approximately 25% of the pages were less than 3 months old. In 2010 and 2011 the age of the pages continued to decrease. In 2011, approximately 45% of the pages were less than one month old. On the Spanish Web, in 2009, approximately 20% of pages had less than 3 months, and 13% were between 12 and 15 months. In 2010 and 2011, the number of pages with less than 3 months increased to 35% and 37%, respectively. On the other hand, we observe that both in 2010 as in 2011 have dropped the pages older than 6 months.

Comparing the age of web pages of the Global and Spanish Web, web pages on the Global Web are updated more frequently, i.e., the age is smaller. In 2011, on the Spanish Web the 37% of the pages were aged less than 3 months versus 45% of those the Global Web. Despite this difference, the age of the Spanish web pages has decreased from 2009 to 2011 and this trend will probably continue in subsequent years.

4) Compression of the content: The compression ratio is a value that represents the relation between the size of the compressed content and the size of the content uncompressed. That is, $CR = 1 - (Size\ Compressed / Size\ Uncompressed)$. At a higher level of compression the content will show more similarity and therefore will often be of lower quality.

Figure 3 shows the obtained results. The results show that on the Global Web in 2009 the compression level was 0.32, and this level grew subsequent years to 0.38 in 2010 and 0.36



(a) Global Web - 2009

(b) Global Web - 2010

(c) Global Web - 2011

(d) Spanish Web - 2009

(e) Spanish Web - 2010

(f) Spanish Web - 2011

Fig. 3: Compression level of contents on Global and Spanish Web

in 2011. With respect to the Spanish Web, we see that in 2009 the average level of compression was 0.43, and that in the next two years, 2010 and 2011, compression levels decreased to values between 0.38 and 0.21. These results indicate that the Spanish Web has dropped its level of compression along the years, and therefore the web pages on Spanish Web have increased and improved their contents. This fact has not occurred on the Global Web, where level of compression is similar along studied years. In summary, the results show that the compression level is higher on the Global Web than on the Spanish Web. This indicates that the contents of web pages on the Global Web, overall, are more repetitive than the Spanish web pages, since the Global web pages have a higher compression level.

5) Title length of web pages: The title of a web page is one of the most important elements in a web page. The use of descriptive titles is important to the Web usability, since it allows to web users to know the topic of the web page. Analysing the results for the Global Web we found that in 2009 the average title length was 7.26 words, in 2010 6.94 and 7.11 words in 2011. From this we can deduce that the value has remained relatively constant at about 7 words. In the results for the Spanish Web, there has been no significant changes from 2009 to 2011. The average length is about 5 words, with pages existing which have a single word or more than 35 words. Both on the Global Web and the Spanish Web, the amount of words in the title has remained relatively constant over the years studied. However, overall on the Global Web the titles of web pages are larger than those of the Spanish Web.

6) Similarity: Indicates the level of similarity, or difference, which have the content of two web pages. Concretely, the element that we have compared is the useful content of each web page, using the same approach that we have described in the Section V-B.

In order to compare the useful content of each web page, we have used a tool implemented by Viliam Holub[4]. This tool divides each document in n tokens, each one of them with a weight. After this, it makes a hash of each of the tokens. Finally, with the weight and the hash of each token, the tool creates a hash for each document of each subset, which “summarizes” its contents.

For our test, we have created a dataset per year, which had 10 random subsets of 10,000 web pages, and we have generated the hash of every page for each year. Then, we compute the Hamming distance among the document signature. The end result, shown in Figure 4, was obtained as the average of the results of each of the 10 subsets for each year.

Figure 4 (a, b, c) shows the results obtained on the Global Web. In 2009, 24% of the Global Web had between 50% and 60% of similarity, more than 30% had between 60% and 70% and 24% had between 70% and 80% of similarity. In 2010 and 2011, the similarity for values between 50% and 60% decreased, but the similarity for values between 60% and 70% increased as did that for values with higher levels of similarity.

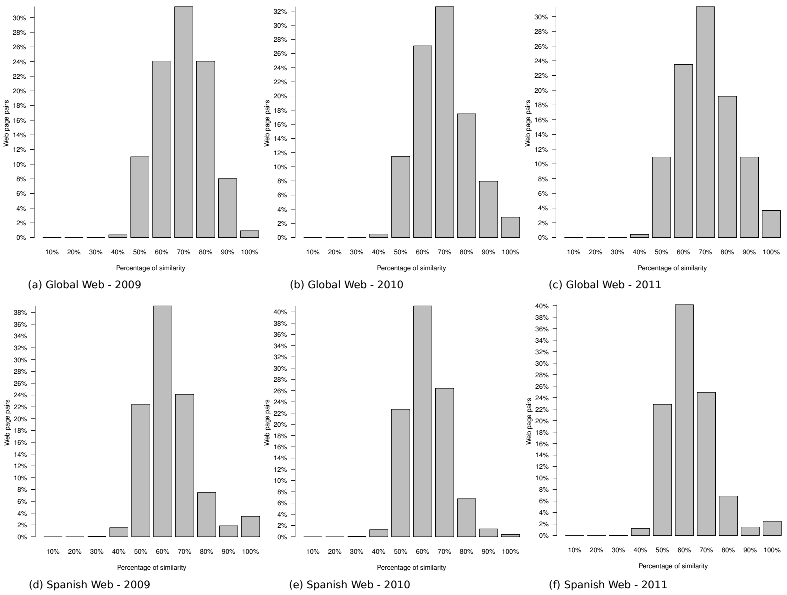
The results about similarity on the Spanish Web are shown in Figure 4 (d, e, f). In summary, from our results we can ascertain that the similarity between pages has remained constant for the 3 years studied. In 2009, approximately 37% of the Spanish Web had between 50% and 60% similarity. In 2010 and 2011, this value increased to 40%. In the 3 years studied, 22% of the Spanish Web had similarity between 60% and 70%.

Based on these results, the Global Web has a higher similarity than the Spanish Web. In the results for 2011, the Global Web has more than 30% of values between 60% and 70% similarity, compared to 25% of the Spanish Web.

These results are consistent with the results of the web page compression showed in Section V-C4, where we could see that the compression level on the Global Web is higher than on the Spanish Web. Therefore, the web pages on the Global Web are more similar than the Spanish web pages.

Comparing our results with those obtained by Cho *et al.* [31], we see that since 2000 the similarity has increased from

[4] http://d3s.mff.cuni.cz/~holub/sw/shash/



(a) Global Web - 2009 (b) Global Web - 2010 (c) Global Web - 2011

(d) Spanish Web - 2009 (e) Spanish Web - 2010 (f) Spanish Web - 2011

Fig. 4: Similarity of the web pages on the Global and Spanish Web

20% or 40% [31] to 50% or 60% in 2011. This increase is quite significant and is an indicator of the evolution of the quality of web content has been declining over the years.

D. Features at Web Site Level

In this level we will explain the main characteristics of a web site, such as the technologies used in the client and server side, and the number of links and their types (static, dynamic, inlink, outlink):

1) Links: The obtained results indicate that on the Global Web in 2009 the average number of links to a web site was 6,502. In 2010 and 2011, the average number of links was 4,639 and 5,773, respectively. Analysing the Spanish Web, these results are inferior in comparison. In 2009, a web site had on average 1,831 links. In 2010, this figure increased by 24.13% to 2,273. In the final year it also grew by 4.31%, to 2,371, compared to 2010.

In summary, the number of links per web site on the Global Web is more than double that of the Spanish Web. This indicates that web sites have a higher overall number of relationships with other websites than with the Spanish Web. However, the evolution of the Spanish Web is greater than the Global Web. The average number of links on a website on the Spanish Web, has increased by 29.49% since 2009, while on the Global Web it has decreased by 11.22% since 2009.

Table II, shows the results regarding inlinks, outlinks, dynamic and static links, and the kind of path, absolute or relative. The results obtained indicate that, in 2009, the inlinks of the Global Web represented 82.35% and the outlinks 17.65%. These results remained similar throughout 2010 and 2011.

Analysing the type of link path, we observe that the relative links dominate the absolute address. In 2009, 60.88% of links were relative versus 29.12% which were absolute. This result was similar in 2010 and 2011, where 63.22% and 36.78% were relative and absolute links, respectively.

	Global Web					
	Inlink	Outlink	Static	Dynamic	Relative	Absolute
2009	82.35%	17.65%	79.47%	20.53%	60.88%	39.12%
2010	84.68%	15.32%	85.45%	14.55%	60.99%	39.01%
2011	84.57%	15.43%	88.20%	11.08%	63.22%	36.78%
	Spanish Web					
	Inlink	Outlink	Static	Dynamic	Relative	Absolute
2009	56.74%	43.26%	71.97%	28.03%	46.98%	53.02%
2010	60.70%	39.30%	77.70%	22.30%	46.65%	53.55%
2011	62.94%	37.06%	76.97%	22.03%	49.14%	50.86%

TABLE II: Links: inlinks, outlinks, static, dynamic, relative and absolute

Analysing the results for dynamic and static links, a domination of static links over dynamic links can be observed. In 2009, the static links were 79.47% as opposed to 20.53% for dynamic links, percentages which increased in 2011, reaching 88.2% for static links and 11.8% for dynamic links. This percentage increased to 88.2% of static links and 11.8% of dynamic links in 2011.

Analysing the results of the Spanish Web, we note that the number of inlinks to web sites has changed significantly. In 2009, 54.74% of links were inlinks and 43.26% were outlinks. In 2011, this number has changed significantly, with inlinks increasing to 62.94% and outlinks decreasing to 37.06%.

Taking into account the path of the link, approximately 50% of the links were relative and the other 50% were absolute links. This number has not changed over the 3 years.

Analysing data on static and dynamic links, we note that the results of each type are very similar for all 3 years. In 2009, 71.97% of links were static and 28.03% dynamic. These data changed in 2011, static links increasing by 5% and dynamic links decreasing by the same amount.

The static link against the dynamic goes slightly against the trend of use of technologies that enable dynamic access to data. This result may be due to the method of coding the URLs, which do not use ?, or that use unknown dynamic extensions.

2) Web forms: As explained in Section III, the server-side Hidden Web is an important part of the Web. In the analysis of the use of web forms, we observed that the Global Web has on average more than 1.1 web forms per web site. This indicates that most web sites use forms to access to certain information, and have, in some cases, more than one. On the Spanish Web this number decreases to 0.4 web forms per web site, i.e., one of two web sites use web forms to access information. In both cases, we found no remarkable changes in the evolution of the use of web forms, over the 3 years studied.

3) Server-side technologies: The results about the Server-Side technologies are shown in Table I. In 2009, the use of client side technologies was widely diversified. The dominant technology was PHP with 25.17%, followed by JSP with 22% CGI 16.18%, SHTML with 17.66% and ASP 15.66%. In 2010 there was a significant change. On one hand, the use of ASP increased to 55.35%, but on the other the use of PHP, CGI and Perl reduced dramatically. This trend was confirmed in 2011, ASP having approximately 50% usage, followed by PHP with 30%.

On Spanish Web, in 2009, 2010 and 2011 the technology most commonly used was PHP with approximately 70% of the sites, followed by ASP with more than 20%, and JSP, with values much lower at approximately 2%. In these 3 years PHP has remained the same, ASP has increased by 2% and JSP has decreased by 1%. The presence of technologies such as CGI and SHTML has decreased since 2009. The results clearly show the difference in the use of these technologies between the Global and the Spanish Web. On the Spanish Web, PHP is the most commonly used technology, whereas on the Global Web, ASP technology is the most popular followed by PHP.

The differences obtained in the results, also appear in the Web of other countries. In the Brazilian Web 73% of the web sites use PHP [14] and in Chile 78% [9]. However, as occurs in our results with the Global Web, in other countries ASP is more used than PHP, for example, in Africa 63% of web sites use ASP [11].

4) Client-side technologies: These are those technologies which allow the creation of dynamic web sites and improve the user experience.

Table I shows the results. On the Global and Spanish Web, the most commonly used technologies are based on JavaScript. On the Global Web, its usage has increased since 2009 with a 93.63% to 97.15% rise in 2011. The second most commonly used technology on the Global Web is Flash, although its presence has reduced since 2009 with a decrease of 6.28% to 2.84% in 2011. On the Spanish Web the predominant language is also JavaScript, which, in 2009, had 70.67% of the sites, followed by Flash with 28.91%. Similarly on the global level, the use of JavaScript in 2011 increased to 77.01% and the use of Flash decreased to 22.17%.

We also note that since 2009 some languages such as VBScript or Tcl have almost disappeared. These results are mainly due to the widespread use of technologies such as AJAX, and the large number of problems of compatibility and security that Flash is currently experiencing.

E. Features at National Web level

In this section we discuss first the software used by web servers and second the number of new domains and deleted domains over the 3 years considered in the study.

1) Web server: A significant characteristic of the national Web is the type of the web servers used. According on the server and its version, the web site uses different web technologies. Also, this allows us to analyze the use of open source solutions (Apache) against shareware products (IIS).

The obtained results show that the most widely used web server is Apache, which in 2009 was present in 65.23% of the servers and in 2011 its presence rose to 70.13%. The next most commonly used server after Apache is Microsoft IIS, which currently has a market share of 26.94%. Other servers that are present in the Spanish Web are Zeus, Nginx or Lotus.

2) New and deleted domains: An important piece of data for search engines and crawlers is the level of growth and the manner of growth in a country. To analyse the amount, we start with the total number of .es domains when the study began: 1,207,832. The following year there was a total of 76,277 new domains and 36,131 deleted domains, which means there were 40,146 more domains than 2009, i.e., an increase of 3.3%. In 2011 this growth was higher, 210,393 new domains were created, which implies a growth of 16.6% from the previous year. To study the growth of the Spanish Web, we have considered that each domain contains an average of 400 web pages per domain [6]. With this data we can say that the growth of the Spanish web is very high.

F. Features at Global Web level

Along this section we explain the software used by web servers on the Global Web.

1) Web server: On the Global Web the most commonly used web server is Apache with 65.4%, 5% less than on the Spanish Web. The next most widely used web server is Microsoft IIS with 18.22%, about 8% less than on the Spanish Web. Another server that has a significant presence on the Web is Nginx with 10% of the market, which on the Spanish Web had less than 2%. Among other servers which have less than 1% presence on the Web are: Tomcat, IBM Servers, Oracle Server, etc.

Both in the study of the Global Web and the Spanish Web, we found that many of the versions used are not recent. Usually, the system administrators tend to be conservative, so it is likely that they do not want to update the web server version quickly, and prefer using older but more stable versions. However, as Rubin *et al.* demonstrate in [32], the use of outdated versions is a potential security problem because they may contain errors established to engage hosted pages and therefore users who visit them.

VI. Conclusions and Future Work

This paper presents a study about the evolution of the Spanish and Global Web at different levels during the years

2009, 2010 and 2011. What differentiates this paper from other related studies is its more detailed approach, its analysis of the evolution regarding time and observations made from the point of view of search engines and crawling systems. Furthermore, it does not focus on one period, but tries to show the evolution of the analysed features.

Within each of the levels we have discussed the different features included in them: HTML tags, stopwords, content size, useful content, URL length, etc.. Among them, we have focused on the size of the content, web technologies, pages age, similarity of content and growth of the web. For each of the features we have tried to give a logical explanation for the evolution of the results and the comparison between Global and Spanish Web.

Future work will focus on processing the Web and creating new datasets, allowing analysis to continue in the forthcoming years. These subsequent studies will allow us to increase our knowledge on how the Web evolves, and also its users and web site creators. The results of the analysis and extracted patterns will help to create and modify the policy of crawling systems.

Acknowledgment

This research was supported by Xunta de Galicia CN2012/211, the Ministry of Education and Science of Spain and FEDER funds of the European Union (Project TIN2009-14203).
The list of domains "es" of 2009, from which we conducted the crawling process, has been provided by the Spanish Business Public Entity Red.es.

References

[1] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages," in *Special interest tracks and posters of the 14th international conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 902–903.

[2] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," in *Proceedings of the 27th International Conference on Very Large Data Bases*, ser. VLDB '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 129–138.

[3] M. K. Bergman, "The Deep Web: Surfacing Hidden Value," *Journal of Electronic Publishing*, vol. 7, no. 1, Aug. 2001.

[4] Z. Gyongyi and H. G. Molina, "Web Spam Taxonomy," in *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, Apr. 2005.

[5] J. P. Kumar and P. Govindarajulu, "Duplicate and near duplicate documents detection: A review," *European Journal of Scientific Research*, vol. 32, pp. 514–527, 2009.

[6] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis, "Characterization of national web domains," *ACM Trans. Internet Technol.*, vol. 7, May 2007.

[7] S. Sanguanpong, P. Piamsa-nga, S. Keretho, Y. Poovarawan, and S. Warangrit, "Measuring and analysis of the thai world wide web," in *Proceeding of the Asia Pacific Advance Network*, 2000, pp. 225–230.

[8] C. C. Ricardo Baeza-Yates and V. Lopez, "Characteristics of the web of spain," 2005.

[9] R. Baeza-Yates and C. Castillo, "Caracterizando la web chilena," in *In Encuentro Chileno de Ciencias de la Computación. Sociedad Chilena de Ciencias de la Computación*, 2000.

[10] V. M. Prieto, M. Álvarez, and F. Cacheda, "Evolucion de la web espanola y sus implicaciones en crawlers y buscadores," in *II Congreso Español de Recuperación de Información*, P. R. Rafael Berlanga, Ed. SciTePress, 2011, pp. 308–315.

[11] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Structural properties of the African Web," in *Proceedings of the eleventh international conference on World Wide Web*. Honolulu, Hawaii, USA: ACM Press, May 2002.

[12] D. Gomes and M. J. Silva, "Characterizing a national community web," *ACM Trans. Internet Technol.*, vol. 5, no. 3, pp. 508–531, Aug. 2005.

[13] J. a. Miranda and D. Gomes, "How are web characteristics evolving?" in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, ser. HT '09. New York, NY, USA: ACM, 2009, pp. 369–370.

[14] M. Modesto, á. Pereira, N. Ziviani, C. Castillo, and R. Baeza-Yates, "Um novo retrato da Web Brasileira," in *Proceedings of XXXII SEMISH*, São Leopoldo, Brazil, 2005, pp. 2005–2017.

[15] E. Efthimiadis and C. Castillo, "Charting the Greek Web," in *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*. Providence, Rhode Island, USA: American Society for Information Science and Technology, Nov. 2004.

[16] G. Grefenstette and J. Nioche, "Estimation of english and non-english language use on the www," in *Proceedings of Content-Based Multimedia Information Access (RIAO)*, Paris, France, 2000, pp. 237–246.

[17] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl, "Who links to whom: Mining linkage between web sites," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, ser. ICDM '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 51–58.

[18] A. B. Downey, "The structural cause of file size distributions," *SIGMETRICS Perform. Eval. Rev.*, vol. 29, no. 1, pp. 328–329, Jun. 2001.

[19] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1-6, pp. 309 – 320, 2000.

[20] B. A. Huberman and L. A. Adamic, *Nature*, vol. 401, no. 6749, p. 131, Sep. 1999.

[21] D. Shestakov, "Sampling the national deep web," in *Proceedings of the 22nd international conference on Database and expert systems applications - Volume Part I*, ser. DEXA'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 331–340.

[22] D. Lewandowski, "A three-year study on the freshness of web search engine databases," *J. Inf. Sci.*, vol. 34, pp. 817–831, December 2008.

[23] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th international conference on World Wide Web*, ser. WWW '03. New York, NY, USA: ACM, 2003, pp. 669–678.

[24] L. Björneborn and P. Ingwersen, "Toward a basic framework for webometrics," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 1216–1227, 2004.

[25] D. Y. Donghua Pan, Shaogang Qiu, "Web page content extraction method based on link density and statistic," in *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference*, 2008, pp. 1–4.

[26] "The w3 consortium the document object model," http://www.w3.org/DOM/, 2011, [Online; accessed 18-February-2011].

[27] N. Shuyo, "Language detection library for java," 2010. [Online]. Available: http://code.google.com/p/language-detection/

[28] A. S. da Silva, E. A. Veloso, P. B. Golghe, B. Ribeiro-Neto, A. H. F. Laender, and N. Ziviani, "Cobweb ? a crawler for the brazilian web," *String Processing and Information Retrieval, International Symposium on*, vol. 0, p. 184, 1999.

[29] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The query-flow graph: model and applications," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 609–618.

[30] T. Suel and J. Yuan, "Compressing the graph structure of the web," in *Proceedings of the Data Compression Conference*, ser. DCC '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 213–.

[31] J. Cho, N. Shivakumar, and H. Garcia-Molina, "Finding replicated web collections," *SIGMOD Rec.*, vol. 29, no. 2, pp. 355–366, May 2000.

[32] A. Rubin and J. Geer, D.E., "A survey of web security," *Computer Journal*, vol. 31, pp. 34–41, 1998.