Orduña Malea, E.; Ortega, JL.; Aguillo, IF. (2014). Influence of language and file type on the web visibility of top European universities. Aslib Journal of Information Management. 66(1):96-116. doi:10.1108/AJIM-02-2013-0018.

PREPRINT: Orduña Malea, E.; Ortega, JL.; Aguillo, IF. (2014). Influence of language and file type on the web visibility of top European universities. Aslib Proceedings. 66(1):96-116.
http://dx.doi.org/10.1108/AJIM-02-2013-0018

Influence of language and file type on the web visibility of top European universities

Abstract

Purpose

The main objective is to detect whether both file type (a set of rich and web files) and language (English, Spanish, German, French and Italian) influence the web visibility of European universities.

Design/Methodology/Approach

A webometrics analysis of the top 200 European universities (as ranked in the Ranking web of World Universities) was carried out by a manual query for each official URL identified by using the Google search engine (April 2012). A correlation analysis between visibility and file format page count is offered according to language. Finally, a prediction of visibility is shown by using the SMOreg function.

Findings

The results indicate that Spanish and English are the languages that correlate most highly with web visibility. This correlation becomes greater -though moderate- when considering only PDF files.

Research Limitations/Implications

The results are limited due to the low correlation between overall page count and visibility. The lack of an accurate search engine that would assist in link counting procedures makes this process difficult.

Originality/Value

An observed increase in correlation -although moderate- while analysing PDF files (in English and Spanish) is considered to be meaningful. This may indirectly confirm that specific file formats and languages generate different web visibility behaviour on European university websites.

Keywords

Webometrics, European universities, Language metrics, Web visibility, Rich files, Web files.

1. Introduction

Webometrics is an emergent discipline which utilises quantitative methods to describe, on the one hand, the communication processes, contents and consumption thereof on the Internet and, on the other hand, the structures, technologies and services used (Aguillo and Granadino, 2006).

Academic activities constitute one single, yet important, subset of these communication processes due to the high correlation of web impact and visibility indicators with some bibliometric measures (Thelwall, 2008a), that existed long before Altmetrics was known in its current form (Priem *et al.*, 2010).

Among these web indicators, Web Impact Factor (WIF) merits particular attention. Since it was introduced by Ingwersen (1998), this indicator has been used widely in webometric analysis, and consists of dividing the total number of external links that a web domain receives (called web visibility) by the total number of web files stored on the same web domain (called page count). Therefore, this indicator reflects the philosophy of Journal Impact Factor, considering links as an expression of citation, and page count as the means of production.

Notwithstanding, at present its use is not advisable due to well-known mathematical artefacts (Aguillo and Granadino, 2006), especially in the overall analysis of universities (Orduña-Malea *et al.*, 2010). The reason for this is that, statistically, larger page count produces greater visibility, so that both small and big websites can have the same WIF while their corresponding performance may be critically different.

In order to fix some of the inherent shortcomings of this indicator, the Webometric Ranking (WR) was developed to compile the Ranking Web of World Universities (Aguillo *et al.*, 2008). The main purpose of this new indicator was to split the page count variable (denominator) by diversifying the nature of file types considered (especially rich files).

The reason for this approach is that the web contains a wide and diverse range of scientific material (Lawrence and Giles, 1999), which is more evident on academic web spaces like universities (Orduña-Malea and Ontalba-Ruipérez, 2013), where other offline activities such as teaching, transfer and administrative issues are reflected online.

Consequently, page count is a very complex variable. Since it is composed of all files shared online within a website domain, these files can therefore be classified under different perspectives, mainly by nature (research, teaching, etc.), intellectual format (article, book, conference presentation, database, etc.) or file type (PDF, HTML, PPT, etc.).

Whereas the analysis of page count as a whole has been of clear interest in the webometric literature (Thelwall, 2004), the study of the performance of specific file types has been uneven. Under the perspective of webometrics, the file types may be classified in the following categories:

Rich files: although in the web industry "rich" files tend to be video or audio based documents, within the scope of webometrics this nomenclature traditionally refers to content-oriented files like DOC, PPT, and PDF formats (Aguillo *et al.*, 2006). The importance of these files is that they are generally assumed to contain the results of intellectual endeavours such as scientific articles, teaching support material or conference presentations, among others. These rich files have been treated as a proxy of academic activity for institutions (by means of the WR indicator), and for impact assessment of research, where the integrated online impact indicator is particularly noteworthy (Kousha *et al.*, 2010).

Web files: this category comprises all documents created by a web mark-up language, among which static (HTML, XML, etc.) and dynamic web files (PHP, ASP, etc.) may be distinguished. The main characteristic of these documents is that they are created to be read primarily by web browsers, and that provides different content for the reader and the machine. An extensive study of web files within the scope of webometrics has not yet been done.

Multimedia files: video, audio, and graphics belong to this file category. The principal attribute of these documents is that they are not text-oriented, so that search engines do not index the real content but the textual metadata associated with each file. Therefore, the webometric research carried out on these files have been either purely descriptive (Orduña-Malea, 2012) or based on embedded metadata (tags, comments, etc.), especially on some widely used file-sharing platforms such as Flickr (Angus *et al.*, 2008) or Youtube (Kousha *et al.*, 2012), among others.

The diverse nature and purpose of these categories of file types may lead them to attract external inlinks differently (webmasters may decide to create a link to these files or not), and the more quantity of external inlinks received by websites will provide them higher web visibility.

Although web visibility indicator does not take into account whether the links are automatically created by machines (which are then considered as spam) or intellectually generated by a human (which are then considered as mentions), only the latter are of interest for webometric research because their meaning is close to the concept of citation.

The motivations for human link creation – even within such specific and controlled spaces as universities – are complex (Seeber *et al.*, 2012), and professional, research-oriented and informative issues are the main motivations for link creation among university websites (Bar-Ilan, 2005; Wilkinson *et al.*, 2003).

Assuming that these issues lead to the creation of preferred file types (for example a PDF is a file type commonly used to disseminate research content), it may be possible to determine their effect on web visibility.

The possible dependence of web visibility on file types in the context of university websites could provide an insight into the files that have the greatest influence on the visibility of universities on the web (and, indirectly, on the performance of institutions on the WR).

Moreover, rich and web files, by virtue of being more focused on textual content, are also prone to being strongly influenced by the language in which the document is written, which becomes another external variable that may also influence link attraction, and therefore web visibility.

In this sense, English, as the globally accepted international language of science (Garfield, 1967), may play an important role in the web visibility of online resources belonging to academic environments.

Hence, the influence of language on the web visibility of universities – depending on file format – becomes a matter of great interest, even more so in a diverse language environment like Europe, where institutions tend to make use of different languages to communicate with their users (generally English and local).

This means that specific file types written in specific languages may be concentrating the majority of inlinks received by university websites. Consequently, those universities creating such files may have an advantage in ranking positions. Obviously, the existence of such specific file types in large quantities may be reflecting some research activities due to the correlation between web visibility and some bibliometric measures, as commented on previously.

Although there are research fronts focused on the study of languages on the web (discussed in the Related work section), they are primarily focused on the metrics of languages and the influence of cultural and linguistic patterns on the generation of links between institutions. However, a lack of studies centred on analysing the effect of language on university web visibility has been detected, and fewer that analyse this effect according to file type.

2. Objectives

The main objective of this study is to answer the following research question: does file type or language (especially English) influence the visibility of European universities?

In these terms, the secondary objectives are the following:

- To determine page count distribution according to the different file formats and languages.
- To ascertain the proportion of file formats to overall page count for each university.
- To identify possible anomalies in page count calculations by search engines.
- To analyse the correlation between page count and visibility according to file type and language.
- To predict visibility, if possible, according to rich and web files in different languages, by means of learning models applied to regression calculation.

3. Related work

The main research activities related to the study of language usage on the web from a webometrics point of view is offered below.

Language usage metrics

Estimating the extent of language use on the Internet is a sub-discipline that has commonly been treated under different and complementary approaches (Crystal, 2001). A complete taxonomy, grouping

together various types of indicators applicable to the area of language metrics, was proposed by Gerrand (2007), who distinguishes "user activity" (actual use of a language on the Internet), "user profile" (number of active Internet users in each language group), "web presence" (number of web pages written in each language group), and a diversity index (statistical measurement which can be applied to all previous indicators).

As regards "user profile" indicator, the Internet World Stats (http://www.internetworldstats.com) platform should be mentioned, which provides metrics according to country.

The "web presence" and "user profile" measurements are characterised by two main methods: direct language analysis, and the use of search engines.

With respect to the direct language analysis method, the Babel Project (http://alis.isoc.org) and the OCLC reports (Lavoie and O'Neill, 1999; O'Neill *et al.*, 2003) are examples of studies in which direct analysis (using language detection software) of randomly addressed websites has been employed to produce estimates of web presence.

The Funredes project constitutes an example of search engines use to estimate the number of web pages in different languages, using Google as a real database of term occurrences (Pimienta *et al.*, 2009).

The "user profile" and "web presence" methods estimate different aspects of language use, although web presence seems to be more accurate for estimating actual language use in cyberspace. In any case, these two procedures for measuring web presence hold some limitations.

Direct language analysis depends on the range of the language detection software (it should be able to recognise all computer-mediated written languages in the world), whereas the use of commercial search engines depends on coverage (web indexed), suitable query commands and accuracy in SERP (Search Engine Results Page) counts, and different issues related to specific languages that search engines have to face (Moukdad and Cui, 2005; Lazarinis, 2007).

Moreover, Lewandowski (2008) demonstrated that features such as language restrictions do not work properly in some major search engines, which means that results for languages different from the interface language received a lower ranking (the extent of this effect on web visibility has not yet been determined).

Likewise, Lewandowski hypothesises that search engines "do not use static language detection, but instead use graded language detection, in which a certain probability that a document contains a specific language is assigned to each document. This could mean that a document that includes content in different languages is assigned to more than one language, but with a different percentage for each".

Linguistic influence on link generation among universities

Link analysis is a well-studied sub-discipline within webometrics (Thelwall, 2004), although the influence of linguistic and cultural aspects on linking between universities has received limited attention.

Thelwall and Tang (2003) analyse linguistic factors in web linking as part of a study comparing Mainland China and Taiwan universities. The authors found no evidence that English was the language of choice for international link pages, despite it was a widely used language in both academic systems.

Likewise, Vaughan (2006) examines how linguistic and cultural factors affect university relationships analysing the Canadian university system, concluding that views on French Canada are based more on linguistic or cultural difference than geographical location.

Thelwall *et al.* (2003) analyse the 16 largest EU countries using the Altavista search engine in order to determine whether there was evidence that "English is the standard language in the EU for the relatively informal melange of scholarly communication represented by web links". The authors found a clear predominance of English in the European academic context (accounting for 56% of all pages),

concluding that English is a standard web language for linking throughout the EU. Similar results were achieved by Ortega (2007). Notwithstanding, the majority of Western European academic websites are international and multilingual in character, with English and national languages operating in tandem throughout.

In any case, some limitations affect the large amount of interconnectivity between university websites in different countries and languages without there being a high underlying degree of international awareness among them (Thelwall *et al.*, 2003), such as the existence of mirror sites of pages hosted in a different country, or individual large collections of international links, among others.

3. Method

The sample of universities is composed of the top 200 European universities as ranked in the January 2012 edition of the Ranking Web of World Universities (http://www.webometrics.info).

For each university, the official URL was noted. After that, a manual query was performed for each URL using Google search engine, which is recommended for webometric tasks when consistent hit counts are needed (Thelwall, 2008). This query consisted of measuring page count filtered by format and language. All formats, aggregations and languages considered are displayed in Table 1.

Table 1. File formats and languages

The file type selection was conditioned by Google's advanced search option, and comprises the main rich files (DOC, PPT, and PDF), static (HTM and HTML) and dynamic web files (ASP and PHP). As for language, the most 5 widely spoken in Western Europe were selected (English, Spanish, French, German and Italian). Additionally, queries without language restrictions (labelled "all") were performed.

The file formats were filtered by using the "filetype" command whereas the language was selected through the advanced search features of Google. For this reason, two different queries were performed to retrieve static web files (HTM and HTML), although there is no difference between them. Later, these two files were merged in the category "static web files".

All the queries (200 universities x 7 file types x 5 languages = 7,000; additional queries without language and file type constraints were performed) were manually carried out in the first week of April 2012 from the same IP address (158.42.48.24) to avoid differences in data collection due to geographical reasons.

The hit count estimates (a number near the top of the results page estimating the total number of results available to the search engine) for the first SERP of each query was recorded as the page count indicator. Google's website IP address was not monitored ("google.com" was used). The differences among datacentres for hit count estimates are not critical in this study because web domains are not compared to each other.

The language of the interface was English. In this case, the effect identified by Lewandowski (2008) had no influence because it has implications primarily for the ranking of the results but not the number of results. Moreover, accuracy in terms of language identification, at this stage of research, was assumed to be correct.

Since Google does not provide accurate external inlinks for entire websites, the web visibility indicator was performed using the API of Ahrefs (http://ahrefs.com).

After that, all the gathered data were exported into a spreadsheet to be statistically analysed. The XLstat 7.5.2 suite was used as a complement to perform advanced analysis:

Correlation analysis: web visibility was correlated against page count data obtained for each file type in each different language in order to find any possible relationship. Since web data presents a skewed distribution, Spearman was applied in all calculations.

Regression analysis was performed in order to model the relationship between rich and web files.

Principal Component Analysis (PCA) was applied in order to complement correlation analysis by finding causes that explain the variability of the indicators applied to the sample. The Pearson(n) PCA with varimax rotation was applied.

Finally, in order to extend the regression analysis, web mining techniques (using machine learning models) were implemented. To this end, Weka 3.6.7 application (http://www.cs.waikato.ac.nz/ml/weka) was used with the aim of testing visibility indicator predictions according to the different formats.

The SMOreg function was selected as the classifier. This model implements the Support Vector Machine (SVM; a specific learning model) for regression, and was used to generate different prediction models to determine the influence of formats and languages (dependent variables) on visibility (independent variable).

4. Results

4.1 Descriptive analysis

Countries

The 200 universities of the sample represent 25 different countries, where we can highlight the presence of Germany (43 universities), UK (29), Spain (24) and Italy (14). The presence of the languages of these 4 countries is even greater due to the fact that they are spoken in different countries (German is also present in Austria, English in Ireland, and French in some universities in Belgium).

The total number of files (obtained from the sum of each file type considered) according to country is shown in Table 2, where Spain (23,915,449), Germany (20,524,903) and Italy (17,805,310) hold the first three positions. Surprisingly, the United Kingdom only achieves a total of 8,212,303 files. Nonetheless, these global figures should be contextualized according to the performance in each specific file type. For example, the elevated results of Germany and Spain (for PDF, SWP and DWP files) explains the overall performance of these countries.

Table 2. Distribution of files by country

The data also confirms the preponderance of web files within the academic websites considered. All 25 countries surpass 75% in terms of web files (static and dynamic), and 9 surpass 90%. Exceptional cases are Israel (97.18%) and Croatia (96.52%), although the low number of observations (only 1 for Croatia and 4 for Israel) does not make this data representative. In any case, the countries with the highest number of universities (Germany, UK and Spain) exhibit elevated web file percentages (85.55%, 82.07% and 91.26% respectively).

Web files are grouped into static and dynamic files, so that the implementation level of each one by country can be checked in Table 2 as well. In this sense, and considering only countries with a high representation in the sample (at least 10 universities), only the Netherlands shows a balanced distribution (SWP: 41.43%; DWP: 44.08%), whereas Germany (SWP: 52.14%; DWP: 33.41%) and the United Kingdom (SWP: 48.64%; DWP: 33.43%) show a stronger presence of static files, and Spain

(SWP: 39.18%; DWP: 52.08%) and Italy (SWP: 21.72%; DWP: 66.47%) a stronger presence of dynamic files.

As regards rich files, 12 of 25 countries surpass 10% in terms of PDF files whereas DOC files are less used; only Hungary (5.02%) is worth noting in global percentage terms. The use of PPTs is scarce (any country achieves 1%), United Kingdom is the country with more PPT files detected (75,540), constituting only the 0.92% of all their gathered files.

Languages and file formats

The distribution of the 5 languages analysed according to the 7 file types considered is shown in Table 3.

Table 3. Distribution of page count according to different files and languages

The distribution observed for the Spanish, French, German and Italian languages is close to that observed for Spain, France, Germany and Italy in Table 2 (a stronger presence of web files with respect rich files, and a higher use of PDF than the remaining rich files). Additionally, a general predominance of PHP over ASP is also detected for all languages.

However, the number of PHP files in English is unexpectedly high (39,493,910 files). One possible reason is that English is used as a second language in most of the non-English universities considered. This could imply that percentages obtained for the Spanish, French, German and Italian languages correspond fundamentally with results obtained for the universities of their respective countries, whereas this does not work for English (as an international language).

Despite this, the percentages of PDFs, DOCs and even static web files are close to those obtained in Table 2 for UK; only PHP files distort the data. This effect could be thus attributed to the use of English commands in web files (these files use English terms which can add a bias in language detection filtering). This issue will be commented on later.

Universities

In Table 4 the URLs with the highest page count according to both file type and language are presented. For each URL page count, their percentage in relation to overall count and file count is also provided. For example, a query retrieved 125,000 PDF files in Spanish for “ucm.es”, which correspond to 91.91% of all PDF files, and to 4.94% of all files within “ucm.es”.

Table 4. Top universities according to file type and language

Otherwise, we can observe some inconsistencies in Table 4, represented by percentages higher than 100% in some URLs (“epfl.ch”, “uni-giessen.de”, “uni-regensburg.de” and “bath.ac.uk”). These anomalies are commented on and discussed in the following section.

4.2 Anomalies

An “anomaly” arises when the number of pages for all five considered languages (“sum”) is higher than the overall page count of a website (“all”). Thus, the error rate is calculated as the difference between “all” and “sum” data as a percentage, indicating if this is lower than 5%, between 5% and 50%, and higher than 50%.

Table 5 includes the number of URLs in which an anomaly in the page count has been detected.

Table 5. Anomalies per file type and country

A total of 103 URLs (51.5% of the sample) have an anomaly when counting PPT files, although only 3 of these URLs generate an error rate higher than 50%. On the other hand, 102 URLs generate an anomaly with respect to PHP files, and 63 of these (61.76%) represent error rates higher than 50%. The inaccuracy with respect to web files (both static and dynamic) is clearly shown as well.
Table 5 additionally adds country filtering in the detection of page count anomalies. For example, 11 Spanish universities present error rates for PHP files (what constitutes 45.83% of all Spanish universities), whereas the same anomaly is detected for 36 Germany universities (83.72%) and 14 British universities (48.28%).

Although each country shows a different pattern, particular attention should be paid to the elevated error rate in Germany (especially office files) and UK (especially ASP files). PPT files are also problematic both in UK (68.97% of universities) and Italy (92.86%).

The anomalies detected present similarities with the results previously obtained in Table 2. A possible explanation for this phenomenon (apart from some search engine inaccuracies) is the multi-language effect. A specific file may be written in different languages. For example an academic paper could originally be written in German, but may contain abstract and bibliographic references in English.

In these cases, the same file will be retrieved in different language-filtered searches, so the aggregation of files filtered by the five analysed languages could exceed the overall count: some files are being counted more than once.

4.3 Correlation analysis

Table 6 contains the correlation factors between different file types (total and English). Data about count page and web visibility (measured by Ahrefs) are provided as well.

Table 6. Correlation between file types (All and English language)

The high correlation between office files (DOC and PPT) and the moderate correlation of PDF with HTML (r=.56) and PPT (r=.49) are worth noting, as well as the low correlation between visibility and all other variables, PDF (r= .42) being the strongest value.

English results are even worse, PDFs again standing out (r= .34). The correlation values for PDFs in the remaining languages is of particular interest: Spanish (r= .34); Italian (r= .29): French (r= .27); and German (r= .21).

The low correlation of visibility could be explained by the use of Ahrefs as a data source (the correlation between size and visibility is r= .37). In any case, the higher correlation achieved by PDFs should be further analysed.

If PPT and DOC file types are aggregated (labelled OF), the results achieved show again a lower correlation with visibility (r=.24) if compared with that achieved by PDF files, more evidenced in the case of English data (r=.15).

These results indicate that PDFs present unexpectedly different behaviour to office files. In order to test this, a PCA was performed (Fig 1). Since the PCA converts the original set of observations of variables (possible correlated) into a set of values of linearly uncorrelated variables (named principal components), this different performance between PDF and office files can be further analysed.

PREPRINT: Orduña Malea, E.; Ortega, JL.; Aguillo, IF. (2014). Influence of language and file type on the web visibility of top European universities. Aslib Proceedings. 66(1):96-116.
http://dx.doi.org/10.1108/AJIM-02-2013-0018

Figure 1. PCA of the usage percentage of different file formats

The proportion of the 2 dimensions (explained variance) is good (71%). The variables HTML, PDF and DWP correlate with the first component in the vertical axis, and the variables PPT and DOC with the horizontal axis.

This may be interpreted as meaning that the act of publishing in web formats by universities is not the same as for non-web formats. In light of these results, it seems that the PDF is less "rich" than the PPT and the DOC, which appear to be more local and specific formats.

Despite the behaviour of PDF files, the nature of rich files is theoretically quite different to web files. While the first are primarily dedicated to the publication of finished intellectual content, the latter (especially ASP and PHP) represent the technical platform (whereas HTML supposes an intermediate approach, as content, design and scripts are combined).

Figure 2 shows the dispersion between rich and web files, finding a meaningful correlation (r= .56) although the following URL outliers have been detected ("huji.ac.il", "usc.es", "uni-trier.de", "muni.cz", "ucc.ie"). The analysis was repeated avoiding these 5 URLs, obtaining a slightly higher correlation (r= .60), with the following equation:

Rich file (t) = 6.44 + 0.37*Web file (t); R^2= 0.36;

Figure 2. Dispersion between Rich files and Web files

4.4 Visibility prediction

Finally, all data was exported into the Weka application in order to generate visibility prediction models (by using the SMOreg function). All tests carried out using different page count components are shown in Table 7 (where the 5 outlier URLs identified previously were avoided).

Table 7. Correlation coefficients in different SMOreg models

The idea behind this procedure was to combine different page count aggregations in order to predict total visibility (as obtained by Ahrefs).

Different rows and columns in Table 7 indicate the nature of the page count aggregations, which were tested disregarding language (T), aggregating the five languages analysed (5L) and considering only English (EN). The term "Null" in the last row means that only aggregation of the column must be considered (and vice versa). The term "All" in the 4th row means that all file sizes are considered in the aggregation.

For example, all file types (PDF, DOC, PPT, HTM, HTML, ASP and PHP) have a correlation of r= .29 (disregarding language), r= .30 (aggregating English, German, Italian, French and Spanish), and r= .30 (taking only English results into account).

The results displayed in Table 7 confirm the low correlation between page count and web visibility. Likewise, an increase of correlation between page count components is observed when page count is divided into several components, getting the best result (r= .39) with PDF and SWP files in English (and avoiding DOC and PPT files). Indeed, the results obtained by considering only English are generally better than the aggregation of the 5 languages (especially when PDF is a separate component).

The equation of visibility prediction which achieves best correlation (r= .39) between page count components is shown below:

V (tot) = 0.25 * (normalised) PDF (en) + 0.12 * (normalised) SWP (en) + 0.10;

If only PDF files in the most influential languages (English and Spanish) are considered, the correlation between dependent variables is even slightly better (r=.42):

V (tot) = 0.14 * (normalised) PDF (sp) + 0.35 * (normalised) PDF (en) + 0.08;

5. Discussion

A number of shortcomings detected in the methodology might be taken into account to contextualise adequately the results obtained.

True web index

Although crawlers, individual analytic applications and webmaster tools give a more accurate index count of online files than commercial search engines, they are restricted to the administrators of these websites. For that reason commercial search engines are widely used to gather data especially in those cases when information from the whole web is needed, rather than just from a limited set of websites (Thelwall, 2004).

Nonetheless, despite the advantages of commercial search engines, the use of these tools leads to various shortcomings:

- Instability of their results (Bar-Ilan, 2002; Rousseau, 1999).
- Limitations in automatic language detection for multi-language files (Martins and Silva, 2005).
- Existence of significant international biases in coverage (Vaughan & Thelwall, 2004; Vaughan & Zhang, 2007).
- Existence of pages that are duplicates of each other, inflating hit count estimates (Thelwall, 2008b).
- Elimination of near-duplicate pages, so that the number of pages returned by a search engine may be significantly lower than the actual number of matching pages (Henzinger, 2006).
- Indexation of a small fraction of the web (Lawrence and Giles, 1999).
- The existence of web link spam, which consists of adding redundant links to a web page or creating pages that only contain superfluous links (Araujo and Martinez, 2009). The consequence of this is that - apart from decreasing the quality of search results and increasing the cost of each processed query (Gyongyi and Garcia-Molina, 2005) - the global results may be misrepresented. Some works have detected the effect of this distortion according to different factors such as TLD (top level domain) or language (Ntoulas et al., 2006).
- The rapid and dynamic evolution of the web. Fetterly *et al*. (2003) found that the average pace of change varies widely across TLDs ("com" pages changed substantially faster than "gov" and "edu"). Koehler (2004) also found evidence of a greater persistence of "edu" compared to "com". Cho and Garcia-Molina (2000) also note that pages in the German domain ("de") exhibit a significantly higher degree of change than those in any other domain, which relates to the amount of spam on German pages detected by Ntoulas *et al.* (2006), and the high degree of anomalies shown in this research for German language pages.

Despite the above mentioned limitations regarding the nature of the web and the functionality of the search engines used, the method and sample employed are intended to minimise their effects:

- University websites are less exposed to distortion by link spam than commercial web sites and to changing over time (Payne and Thelwall, 2007).

- The use of search engines and hit count estimates is the best procedure for analysing the sample, taking into account the limitations of personal crawlers and the inaccessibility of some data.
- The limitations of search engines (coverage, instability, duplicates) affect all universities in the sample in the same way, so their effects are minimised to some extent.As regards the anomalies in language detection, those under 5% are assumed to be due to the rounding-off procedures of the search engines; from 5% to 50%, files written in different languages (since it is not possible to manually check each file and determine the weight of each language in it, this is the best possible method); and those over 50%, real anomalies (which only affect dynamic web files, therefore having a limited effect on web visibility prediction and correlation coefficients for the remaining file types).
- The dynamism of the web compels the contextualisation of data as a snapshot of the moment at which they were retrieved.

Sample coverage

The sample of European universities could limit the findings. Considering a higher number of institutions, countries and languages will enrich the results. In any case, the analysis of the top 200 universities in terms of web impact, and the inclusion of the main spoken languages, offers a broad picture of the situation.

Multimedia files and some other rich (DOCX, PPTX, etc.) and web files (especially XML) should be included in future works although their presence now on university websites is still scarce.

6. Conclusions

The general conclusions of this research are the following:

Use of file types

The correlation between content (rich files) and infrastructure (web files) is positive and meaningful (r= .60, avoiding 5 URL outliers), although web files represent an important percentage in the overall page count of the universities. This implies that technological support and infrastructure is unexpectedly much higher than content publication.

If web files are disaggregated, PHP achieves the greatest weigh, followed by HTML. The use of ASP is limited. On the other hand, if rich files are disaggregated, only PDF achieves significant use percentage whereas both DOC and PPT are marginal (only significant for English results).

Otherwise, page count procedure presents some anomalies (the overall page count is sometimes lower than that obtained by adding the results of the 5 different languages analysed). This effect is attributed to:

a) Multi-language property: files containing different languages are detected in different language-filtered queries so they are counted more than one time when aggregating results for the 5 languages.
b) Web file commands: script languages use “English” commands so that these files are over-representing English content in the overall results.

File types and web visibility

PDF is the format that best correlates with visibility whereas DOC and PPT correlate very poorly. This effect indicates that rich files should be divided, considering PDF separately.

Finally, all combinations of rich files used to predict visibility with the SMOreg learning model indicate that PDF should be treated separately whereas DOC and PPT files should be avoided. Moreover, English and Spanish have a positive influence.

File types and languages

Spanish and English are the languages that have the highest correlation with web visibility due to both the high representation of UK and Spain in the sample, and the international character of these languages, especially English (used as a second language on practically all university web platforms).

German, despite its high representation, presents a low correlation with visibility, due to a much higher percentage of anomalies in the search engine results than others. Moreover, the high percentage of spam for this language detected in previous studies may have an influence, which should be analysed in further studies.

Italian maintains an intermediate performance due to its expansion outside Italy, especially for PDF Italian content detected in Swiss and German universities, for example.

The presence of French is minimal (mainly as a consequence of the low coverage of French universities in the sample). Additionally, it should be pointed out that Francophone universities such as École Polytechnique Fédérale de Lausanne and Université de Genève (Switzerland), and the Université de Liège and Université Libre de Bruxelles (Belgium) obtain better results than the French universities analysed.

Final considerations

Considering the limitations described previously, this research finds an increase in correlation – although moderate – between page count and visibility when considering PDF files (in English and Spanish).

This may indirectly confirm that specific file formats and languages are influencing and attracting more external links to European universities than others, therefore generating an advantage in ranking positions, what constitutes an affirmative answer to the research question that constitutes the main objective of this study.

These findings help to a better understanding both of online-content creation patterns within university websites as of variables which affect the construction of their web visibility.

References

Aguillo I.F., Granadino, B., Ortega J.L. and Prieto J.A. (2006), "Scientific research activity and communication measured with cybermetrics indicators", *Journal of the American Society for Information Science and Tecnology*, Vol. 57 No. 10, pp. 1296-1302.

Aguillo, I.F. and Granadino, B. (2006), "Indicadores web para medir la presencia de las universidades en la Red", *Revista de universidad y Sociedad del Conocimiento*, Vol. 3 No. 1, pp. 68-75.

Aguillo, I.F., Ortega, J.L. and Fernández, M. (2008), "Webometric ranking of World universities: introduction, methodology, and future developments", *Higher education in Europe*, Vol. 33 No. 2-3, pp.233-244.

Angus, E., Thelwall, M. and Stuart, D. (2008). "General patterns of tag usage among university groups in Flickr", *Online Information Review*, Vol.32, No. 1, pp. 89-101.

Araujo Serna, L. and Martínez Romo, J. (2009), "Detección de Web Spam basada en la recuperación automática de enlaces", *Procesamiento del lenguaje natural*. No. 42, pp. 39-46.

Bar-Ilan, J. (2002), "Methods for measuring search engine performance over time", *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 4, pp. 308–319.

Bar-Ilan, J. (2005), "What do we know about links and linking? A framework for studying links in academic environments", *Information Processing & Management*, Vol. 41, No. 3, pp. 973–986.

Cho, Y. and García-Molina, H. (2000), "The evolution of the web and implications for an incremental crawler", *Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 200-209.

Crystal, D. (2001), *Language and the Internet*, Cambridge University Press, Cambridge.

Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2003), "A large scale study of the evolution of web pages", *Proceedings of the Twelfth International Conference on World Wide Web*, pp. 669–678.

Garfield, E. (1967), "English - An international language for science?", *Current Contents*, No. 19-20.

Gerrand, P. (2007), "Estimating linguistic diversity on the internet: a taxonomy to avoid pitfalls and paradoxes", *Journal of Computer-Mediated Communication*, Vol. 12 No. 4, pp. 1298-1321.

Gyongyi, Z. and Garcia-Molina, H. (2005), "Web Spam Taxonomy". In *First International Workshop on Adversarial Information Retrieval on the Web* (AIRWeb 2005), May 10-14, 2005, Chiba, Japan.

Henzinger, M. (2006), "Finding near-duplicate Web pages: A large-scale evaluation of algorithms", In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.284–291. New York: ACM Press.

Ingwersen, P. (1998), "The calculation of Web Impact Factors". *Journal of Documentation*, Vol. 54 No. 2, pp. 236-243.

Koehler, W. (2004), "A longitudinal study of Web pages continued a consideration of document persistence", *Information Research*, Vol. 9, No. 2.

Kousha, K., Thelwall, M. and Abdoli, M. (2012), "The role of online videos in research communication: a content analysis of YouTube videos cited in academic publications", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 9, pp. 1710-1727.

Kousha, K., Thelwall, M. and Rezaie, S. (2010), "Using the web for research evaluation: the integrated online impact indicator", *Journal of informetrics*, Vol. 4, No.1, pp. 124-135.

Lavoie, B. F. and O'Neill, E. T. (1999), "How 'World Wide' is the Web? Trends in the internationalization of web sites". *Annual Review of OCLC Research 1999*. Retrieved November 27, 2012 from http://worldcat.org/arcviewer/1/OCC/2003/06/11/0000003496/viewer/file1.html

Lawrence, S. and Giles L. (1999), "Accesibility of information on the Web", *Nature*, Vol. 400, pp. 107-109.

Lazarinis, F. (2007), "Web retrieval systems and the Greek language: do they have an understanding", *Journal of information science*, Vol. 33 No. 5, pp. 622-636.

Lewandowski, D. (2008), "Problems with the use of web search engines to find results in foreign languages", *Online information review*, Vol. 32, No. 5, pp. 668-672.

Martins, B. and Silva, M.J. (2005), "Language identification in web pages", *Proceedings of the ACM Symposium of applied computing*. Santa Fe, NM, pp. 764-768.

Moukdad, H. and Cui H. (2005), "How do search engines handle Chinese queries", *Webology*, Vol. 2 No. 3, pp. 17.

Ntoulas, A., Najork, M., Manasse, M. and Fetterly, D. (2006), "Detecting spam web pages through content analysis", *Proceedings of the 15th international conference on World Wide Web*, pp. 83-92

O'Neill, E. T., Lavoie, B. F. and Bennett, R. (2003), "Trends in the evolution of the public Web: 1998–2002", *D-Lib Magazine*, Vol. 9 No. 4, available at: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html (accessed 11 February 2013).

Orduña-Malea, E. (2012). Graphic, multimedia, and blog-content presence in the Spanish academic web-space. *Cybermetrics*, Vol. 16, available at: http://cybermetrics.cindoc.csic.es/articles/v16i1p3.pdf (accessed 11 February 2013).

Orduña-Malea, E. and Ontalba-Ruipérez, J-A. (2013), "Proposal for a multilevel university cybermetric analysis model", *Scientometrics*, Vol. 95, No. 3, pp. 863-884.

Orduña-Malea, E., Serrano-Cobos, J., Ontalba-Ruipérez, J-A. and Lloret-Romero, N. (2010), "Presencia y visibilidad web de las universidades públicas españolas". *Revista española de documentación científica*, Vol. 33, No. 2, pp. 246-278.

Ortega, J.L. (2007), *Visualización de la Web universitaria Europea: análisis cuantitativo de enlaces a través de técnicas cibermétricas*, Universidad Carlos III de Madrid, Madrid

Payne, N. and Thelwall, M. (2007), "A longitudinal study of academic webs: growth and stabilization", *Scientometrics*, Vol. 71, No. 3, pp. 523-539.

Pimienta, D., Prado, D. and Blanco, A. (2009), *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives*, Unesco, Paris.

Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010), "*Altmetrics: a manifesto*", available at: http://altmetrics.org/manifesto (accessed 11 February 2013).

Rousseau, R. (1999), "Time evolution of the number of hits in keyword searches on the Internet", *Post Conference Seminar – Cybermetrics'99 at the Seventh International Conference on Scientometrics and Informetrics*, July 9 1999, Colima, Mexico. Available at: *http://www.cindoc.csic.es/cybermetrics/cybermetrics99.html* (accessed 11 February 2013).

Seeber, M., Lepori, B., Lomi, A., Aguillo, I. and Barberio, V. (2012), "Factors affecting web links between European higher education institutions", *Journal of informetrics*, Vol.6, pp. 435–447.

Thelwall, M. (2004), Link analysis: An information science approach, San Diego: Academic Press.

Thelwall, M. (2008a), "Bibliometrics to webometrics", *Journal of Information Science*, Vol. 34 No. 4, pp. 605-621.

Thelwall, M. (2008b). "Quantitative comparisons of search engine results". *Journal of the American Society for Information Science and Technology*, Vol. 59, No. 11, pp. 1702-1710.

Thelwall, M. and Tang, R. (2003), "Disciplinary and linguistic considerations for academic web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan", *Scientometrics*, Vol. 58, No. 1, pp. 155-181.

Thelwall, M., Tang, R. and Price, L. (2003), "Linguistic patterns of Academic web use in Western Europe", *Scientometrics*, Vol. 56, No. 3, pp.417-432.

Vaughan, L. (2006), "Visualizing linguistic and cultural differences using web co-link data", *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 9, pp. 1178-1193.

Vaughan, L. and Thelwall, M. (2004), "Search engine coverage bias: Evidence and possible causes", *Information Processing & Management*, Vol. 40, No. 4, pp. 693–707.

Vaughan, L. and Zhang, Y. (2007), "Equal representation by search engines? A comparison of Websites across countries and domains", *Journal of Computer-Mediated Communication*, Vol. 12, No. 3.

Wilkinson, D., Harries, G., Thelwall, M. and Price, L. (2003), "Motivations for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication", *Journal of information science*, Vol. 29, No. 1, pp. 49-56.