PREPRINT VERSION - PLEASE CITE AS:
Ortega Jose L., Orduña-Malea, E., & Aguillo, Isidro F (2014). Are web mentions accurate substitutes for inlinks for Spanish universities?. Online Information Review, 38(1), 59-77.
http://www.emeraldinsight.com/doi/full/10.1108/OIR-10-2012-0189

Are web mentions accurate substitutes for inlinks for Spanish universities?

José Luis Ortega
Vice-presidency for Scientific and Technological Research,
Spanish National Research Council, Madrid, Spain

Enrique Orduña-Malea
EC3 Research Group and Institute of Design and Manufacturing,
Polytechnic University of Valencia, Valencia, Spain, and

Isidro F. Aguillo
Institute of Public Goods and Policies, Spanish National Research Council,
Madrid, Spain

Abstract

Purpose

Title and URL mentions have recently been proposed as web visibility indicators instead of inlink counts. The objective of this study is to determine the accuracy of these alternative web mention indicators in the Spanish academic system, taking into account their complexity (multi-domains) and diversity (different official languages).

Design/Methodology/Approach

Inlinks, Title and URL mentions from 76 Spanish universities were manually extracted from the main search engines (*Google*, *Google Scholar*, *Yahoo!*, *Bing* and *Exalead*). Several statistical methods, such as correlation, difference tests and regression models, were used.

Findings

Web mentions, despite some limitations, can be used as substitutes for inlinks in the Spanish academic system, although these indicators are more likely to be influenced by the environment (language, web domain policy, etc.) than inlinks.

Research Limitations/Implications

Title mentions provide unstable results caused by the multiple name variants which an institution can present (such as acronyms and other language versions). URL mentions are more stable, but they may present atypical points due to some shortcomings, the effect of which is that URL mentions do not have the same meaning as inlinks.

Practical implications

Web mentions should be used with caution and after a cleaning-up process. Moreover, these counts do not necessary signify connectivity, so their use in global web analysis should be limited.

Originality/Value

Web mentions have previously been used in some specific academic systems (US, UK and China), but this study analyses, in depth and for the first time, an entire non-English speaking European country (Spain), with complex academic web behaviour, which helps to better explain previous web mention results.

Keywords

Webometrics, Link analysis, Web mentions, Title mentions, Search engines, Universities, Spain.

1. Introduction

University websites have gradually become complex systems of dynamic information where both institutions and services are linked and potentially accessible from a general URL to potential users such as students, teachers, researchers, companies, and so on (Orduña-Malea and Ontalba-Ruipérez, 2012).

The quantitative analysis of all the data contained within these online systems could bring to light information unobtainable through other research methods such as bibliometrics (Aguillo 2009), providing a complementary understanding of general university performance. The design of university web rankings constitutes an example of the applied use of web data in the creation of university evaluation tools[1][2].

The discipline of cybermetrics provides the theoretical basis and methodology necessary for a quantitative analysis of the information contained on university websites (Björneborn and Ingwersen, 2004), but the heavy reliance of this discipline on search engines to collect web data means that the most accurate and available web indicators (especially inlink counts) and appropriate procedures (Thelwall and Sud, 2012) for obtaining these data have to be reviewed periodically.

Recently, due to an important change in the search engine market (commented on later), Title and URL mentions have been proposed as web visibility indicators instead of the traditional inlink counts for specific academic environments (China, UK and US). The main objective of this study is to determine whether these alternative web mention indicators could be generalised to other university systems (especially in the Spanish academic web system) and consequently be employed, particularly in global university web rankings.

In order to address this question, a description of a key event in the search market with several implications for webometric methodologies (agreement between *Microsoft* and *Yahoo!*), and the main actions carried out by the scientific community to avoid them (mainly the proposal of alternative indicators and sources) are provided below. After this, a gap in research (the widespread use of alternative web mention indicators and their application in global web rankings) is identified and commented on. Finally, and in view of this gap, specific research objectives are set out.

1.1. *Web search market changes*

Recently the web search market has undergone important changes that have affected the availability of data on linking relationships between web sites and domains. Previously, the most reliable sources (and those with the largest coverage for extracting data on linking patterns) were the *Yahoo! Search* and *Yahoo! Search Explorer* (YSE) databases. However, in July 2009, *Microsoft* and *Yahoo!* announced a commercial and technological agreement in which, among other things, *Bing* would be the exclusive search engine for both companies (*The Washington Post*, 2009). Since Bing did not support the "link:" and "linkdomain:" advanced query operators (Seidman, 2007), the possibility of obtaining this type of information was jeopardised.

Empirical testing showed that this integration took place around October 2011, while the "link:" and "linkdomain:" operators gradually disappeared from each local search portal. In November 2011, therefore, YSE turned off the service permanently, and the main source for large-scale selective link information thus disappeared.

Today, the only general search engine that supports link searches (selecting source and target) is *Exalead,* but its coverage is not only relatively limited but also has a strong geographical bias (Orduña-Malea *et al.*, 2010). Other link information services are *Open Site Explorer*[3], *Majestic SEO*[4] and *Ahrefs*[5], but these services do not allow the source of the inbound links to be distinguished, and the coverage is weak if compared to general search engines (Orduña-Malea, 2012).

A solution to this problem is to use alternative ways, which do not involve link operators, to find relationships between web sites or domains. Two web mention types have been proposed: Title mentions and URL mentions.

Title mention refers to the number of times that the title of a document, the name of an institution, topic, object or person appears in the results of a search engine query. One of the first approaches to Title mentions was proposed by Cronin *et al.* (1998) through the exploration of different ways to invoke scholars on the Web. But a more formal use was adopted by Vaughan and Shaw (2003) to establish a relationship between the number of times the title of scientific articles appear in search engines and their citations in the *Social Science Citation Index*, and by Kretschmer and Aguillo (2004) to identify networks of authors on the Web. This citation type is useful to identify relationships in documents where it is not possible to extract links, such as online presentations (Thelwall and Kousha, 2008) and *Google Books* (Kousha and Thelwall, 2009), as well as to sound out public perception of several organisations through the co-occurrence of their names on the Web (Vaughan and Young, 2010). Likewise, Vaughan and Romero-Frias (2012), who used the term "web keywords" to refer to Title mentions, analysed the occurrence of the name of American companies on the Web and studied them in relation to business indicators. Their results also suggested that keyword count could replace inlink count as an alternative indicator in a non-academic environment.

On the other hand, URL mention is similar to Title mention, with the difference that the requested string is the web domain (or web address) instead of the name of an organisation. This could be a more precise indicator since it would be closer to the hyperlink concept because the appearance of a URL in a text expresses a transitivity relationship to the referenced source, while the citation of the title of an organisation may be in different contexts, such as acknowledgements, citations, lists, etc.

The URL mention has been used less. The work of Zhang (2006), who proposed its use to count citations between articles published in open access journals, may be highlighted, while Stuart and Thelwall (2006) used it to extract triple helix relationships on the Web. Nowadays, it is proposed as a serious alternative to inlinks in view of their disappearance from search engines, as previously commented on.

Thelwall (2011) was the first to compare the performance of URL mentions with inlinks by testing both impact measures in different web domains, finding that URL mentions are less numerous than inlinks, but that their count increases in academic website environments. Thelwall concluded that the low results in URL counts would undermine the effectiveness of link analysis, except in the case of university web studies.

Later, Thelwall and Sud (2011) extended their previous study by adding the organisation Title mention. They employed correlation analyses to test these relationships between indicators and between different search engines, concluding that the high correlations among these three types of web mentions could be used interchangeably for web impact measurements. A subsequent study analysed both binary and weighted link network matrices from these types of web mentions, finding that the best type of data to construct web network diagrams were the filtered URL counts (Thelwall, Sud and Wilkinson, 2012).

Recently, Vaughan (2012) also pointed out the advantages of the *Alexa* "sites linking in" command as an alternative to *Yahoo!* inlink count. Later on, Vaughan and Yang (2012) applied this indicator to analyse two large samples (universities and companies) in two different areas (United States and China), correlating the results obtained with *Yahoo!* inlink data and *Google* URL mention (called URL citation), concluding that both *Alexa* inlink and *Google* URL citation data can replace *Yahoo!* inlink data, and that the former is better than the latter.

1.2. *Research gaps*

Although all the previous works have tested the suitability of URL mentions as a proxy for inlinks, it should be pointed out that all these studies have been performed in specific academic environments, and not in a global arena. Thelwall and Sud (2011) and Thelwall, Sud and Wilkinson (2012) analyse US library and information science departments and UK universities, while Vaughan and Yang (2012) analyse US and China universities (leaving apart business companies).

Notwithstanding, web academic systems are widely diverse, and the assumption that both URL and Title mentions could be proxies should not be generalised a priori to other academic environments, with different web policies and technical infrastructure. In fact, Vaughan and Yang (2012) found a correlation of 0.91 between *Google* URL mentions and *Yahoo!* inlink count for US universities, whereas this correlation drops to 0.70 if Chinese universities are considered. Therefore, the following general research question arises: if the correlation between web mentions and inlinks varies excessively between different academic systems, can web mentions be used accurately in global web analysis?

In this sense, the Spanish system (composed of 76 official universities in 2012) has a specific web environment, as showed recently by Orduña-Malea (in press). Some Spanish universities can be named in different ways because there are four official languages (Castilian, Catalonian, Basque and Galician), and this could influence the correlations between *Yahoo!* inlinks and Title mentions. Regarding URL mentions, the Spanish web does not have a second-level domain for academic institutions, as is the case in the United Kingdom (“ac.uk”), and moreover, a strong multi-domain activity was detected, that is, universities holding more than one valid official web domain. Additionally, the differences between British and Spanish university systems have been previously detected and well described (Thelwall and Aguillo, 2003).

As a consequence of the well-known multi-domain activity in the Spanish system, *Alexa*’s “sites linking in” command should not be employed, because it is applied to only one web domain per university, as showed by Vaughan and Yang (2012), which implies an underrepresentation of real inlinks. Moreover, *Alexa* is based on user panels (a sample of users), and the coverage for Spain is lower than in English speaking countries.

Due to all the reasons stated above, an analysis of URL and Title mentions in the Spanish web system is necessary in order to check their accuracy as link predictors, and thus reinforce (or not) the previous studies in other academic environments.

1.3. *Objectives*

The main objective of this study is to determine whether the use of alternative web mention indicators (Title and URL) is influenced by the diversity of the Spanish academic web system or not. That is, if the use of Title and URL mentions as substitutes for inlinks is accurate for Spanish universities in the same way that it is for other already studied university systems, so that their use in global analysis and rankings may be reinforced.

The specific goals of this research are set out below:

- To analyse the relationship between the different types of web mentions (inlinks, Title mentions and URL mentions) taking into account the different languages and multi-domains of the Spanish universities.
- To determine the extent to which URL mentions and Title mentions could be a replacement for inlink counts.
- To quantify and estimate the number of inlinks that a website receives from URL mentions and Title mentions.
- To explore the advantages and drawbacks of web mentions, and the possible limitation, if any, in employing them as a replacement for inlink count in Spain.

2. Literature review

Cybermetrics has traditionally paid particular attention to the definition of units of measurement and the description and application of web-based indicators. This activity has reflected in diverse projects with European funding, such as the WISER project (Web Indicators for Science, Technology and Innovation Research)[6], with its Indicators Web Portal[7], the EICSTES project (European Indicators, Cyberspace and the Science-Technology-Economy System)[8], and recently, the ACUMEN Project (Academic Careers Understood through Measurement and Norms)[9].

Among the indicators studied in the aforementioned projects, the measurement of mentions is of particular interest due to their accuracy in measuring the impact and popularity of online assets. Among them, hyperlinks have been widely used in the web analysis of university systems because these spaces constitute excellent test beds both for testing the characteristics of links and for studying the relationship between universities.

Characteristics of links

The motivation behind the creation of links is essential to comprehend the nature of web impact. Notwithstanding, these motivations are not easy to define as they cannot be directly related to specific types of relationships (Seeber *et al.*, 2012).

Smith (1999) and Thelwall (2001) outlined motivations for link creation such as referring to educational or informative materials. Thelwall (2002a) also showed that motivations were largely related to the main activities of universities. Wilkinson *et al.* (2003) studied link patterns between UK universities and found that 90% of links were created for scholarly-related activities, and Harries *et al.* (2004) studied links between academic websites in different disciplines (mathematics, physics and sociology), finding differences for each one.

Bar-Ilan (2004) analysed Israeli universities and found motivations for link creation such as signalling the institutional space to which the university belongs or referring to useful information in the same geographical area, amongst others. Later on, Bar-Ilan (2005) found that the main motivations for link creation in Israeli universities were professional and work-related (32%), research-oriented (28%) and informative (14%).

Finally, Seeber *et al.* (2012) analysed factors pertinent to web links within European Higher Education Institutions concluding that, while the presence of a web link cannot be directly related to its underlying motivation, patterns of network ties between universities present statistical properties which reveal new insights on the function and structure of the inter-organizational networks in which these universities are embedded.

Relationship between universities and networks

In Europe, Boudorides *et al.* (1999) and Thelwall *et al.* (2002) were pioneers in visualizing the relationships between European university websites, and Ortega *et al.*

(2008) found that European-level interlinking patterns were set up by the aggregation of national networks, where Germany and UK were dominant.

Thelwall (2002) found that the number of links between pairs of universities in the UK decreased with distance. Later on, Thelwall et al. (2003) found that universities tended mostly to link to countries with a shared language or geographically close. Heimeriks *et al.* (2003) also detected cultural and linguistic patterns by mapping 220 European universities, whereas Heimeriks and Van den Besselaar (2006) found that international linking was also associated with country sizes, whereas Thelwall and Zuccala (2008) detected a dominance of the large richer Western European nations, particularly the UK and Germany. Ortega and Aguillo (2008) found that the Finnish academic web space was isolated from Europe.

Outside Europe, the studies of Israel (Bar-Ilan, 2003), China (Qiu et al, 2004) and Iran (Kousha and Horri, 2004) should be highlighted. South America (Ortega and Aguillo, 2009a), North America (Ortega and Aguillo, 2009b), Canada (Vaughan et al., 2007) and Australia (Smith and Thelwall, 2002) have been also analyzed, while Africa has been studied only partially (Adecannby, 2011).

The scale of academic data gleaned from link analysis has enabled the compilation of university web rankings, where external inlinks constitute a key indicator in their methodology (Aguillo *et al.*, 2005; Aguillo *et al.*, 2008).

3. Method

First, the data gathering process is outlined, and then the statistical analysis is commented on.

3.1. *Data gathering*

Web mention data is extracted from a range of general search engines (*Google*, *Yahoo! Search*, *Bing* and *Exalead*) with the intention of identifying the relationship between these web mention types and the implications they may have for webometric studies in a distinct web environment such as the Spanish academic web. Additionally, *Google Scholar* is used in order to test whether the web mention indicators are more accurate when treating academic content.

The list of Spanish universities (76) with their web domains was compiled in order to obtain their web citations. These data were directly extracted from the search engines in November 2011 to avoid any fluctuation and anomalies in the results. This process was performed before the complete integration of *Yahoo! Search* and *Bing* from the Spanish mirror of *Yahoo! Search*, still operative at the beginning of November 2011.

From each university web domain the following type of web mentions were extracted:

Title mentions:

These are defined as the number of times that the title of a website or the name of an institution is invoked in a search engine minus the Title mentions recorded in their own web domains. For example, to obtain the Title mentions of the *Universidad Complutense of Madrid*, our query would be <*"Universidad Complutense de Madrid" -site:ucm.es*>. This retrieves all the mentions on the Complutense University of Madrid on the web pages indexed by a search engine, excluding the pages hosted in the “ucm.es” sites.

The search services used to obtain this information were *Google, Bing, Yahoo! Search*, *Exalead* and *Google Scholar*.

Due to the fact that some universities have different official names in some of the Spanish official national languages (Catalonian, Basque, Galician, etc.), we calculated the sum of title variants for each university by performing a query for each language variant. A total of 116 mentions were used for the 76 universities. For example:

<"Universidad del Pais Vasco" -site:ehu.es>
<"Euskal Herriko Unibertsitatea" -site:ehu.es>

URL mentions:

These are similar to the Title mentions, the difference being that the URL of the site is used instead of the title. Following the previous example, the URL mention query is <"ucm.es" -site:ucm.es>, which retrieves the number of URL appearances on the pages indexed in the search engines minus the URL mentions of the "ucm.es" sites.

In this case, the same search engines were used: *Google*, *Bing*, *Yahoo! Search*, *Exalead* and *Google Scholar*.

It was found that a few universities have different web domains (i.e., ub.cat, ub.edu, ub.es), so the total number of URL mentions of each domain was aggregated. A total of 145 URL mentions were taken into account. For example:

<"ub.cat" -site:ub.cat -site:ub.edu -site:ub.es>
<"ub.edu" -site:ub.cat -site:ub.edu -site:ub.es>
<"ub.es" -site:ub.cat -site:ub.edu -site:ub.es>

Inlinks

This is the most extended and used web mention type (Aguillo *et al.*, 2006). It is defined as the number of hypertext links that a website or domain receives from all the web pages indexed in a search engine. At the moment of collection, only two important search engines allowed this information this to be obtained (*Yahoo! Search* -and its YSE service- and *Exalead*).

Following the previous example, the query used in both services is "linkdomain:ucm.es –site:ucm.es", which extracts all links that point to the "ucm.es" domain minus their Internal links.

As with the URL mentions, the inlinks of universities with several domains were added, using a total of 145 web domains. For example:

< linkdomain:uib.cat -site:uib.cat -site:uib.es>
<linkdomain:uib.es -site:uib.cat -site:uib.es>

Additionally, for some universities, where both alternative domains and titles were found, all combinations were added, for example:

<"Universidad de Lerida" -site:udl.cat -site:udl.es>
<"Universidad de Lleida" -site:udl.cat -site:udl.es>
<"Universitat de Lleida" -site:udl.cat -site:udl.es>

3.2. *Statistical analysis*

As the Web shows scale-free properties and the distribution of links follows a power law (Barabasi and Albert, 1999), the statistical analysis of these data entails the use of non-parametric statistics and logarithm transformations because the arithmetic mean in these cases is not appropriate due to the skewed distribution of data. The different statistical tests and measures performed are set out below:

Correlation

A correlation coefficient is a dependence measure that allows relationships between variables to be detected. These relationships are always symmetric because the coefficient only measures their reciprocal influence and it does not determine which variable affects the other one. The correlation may reflect intensity and direction. An intense correlation shows a strong relationship between variables and the direction indicates if this is direct (positive) or inverse (negative). Due to the non-parametric behaviour of data, the Spearman correlation was considered.

Differences between samples

The Friedman test (1937) for *n* samples was used to study the statistical differences between the web mention types and the different search engines that provide these data. This is a non-parametric test, similar to the ANOVA parametric test, which detects differences between paired samples. It was used in combination with the *post hoc* Nemenyi test (1963) which points out the samples that differ between themselves.

Regression analysis

A regression model was used in order to address one of our objectives: to quantify and estimate the relationship between the inlinks and the other web mentions (Title and URL mentions). Linear regression allows us to determine whether there is a relationship of dependence between variables and the weight of each variable in the model. Regression goes beyond correlation by adding prediction capabilities and makes it possible to determine whether Title mentions and URL mentions may predict the inlinks that a site receives and to estimate the margin of error of that prediction.

Two assumptions about this model are necessary: the independence of the observations and the normality of the distribution. The first states that none of the observations determines the following one. The second assumption requires the variables to have a normal distribution, whose density function must be symmetric. Given the non-normality of the web data distribution, a possible alternative is the use of non-parametric regression models, although the most immediate alternative is the transformation of the dependent variable (Bland and Altman, 1996). In this case, the variables used in this study have been transformed into logarithms.

All these statistical tests were performed with SPSS 19 and XLStat 2008 statistical packages.

4. Results

Firstly, the distribution of data obtained is described, followed by a description of the correlation and regression model.

4.1. *Distribution of data*

A visual distribution of the data is presented to describe the singularities of each web mention type, which allow us to comment on some drawbacks in the data extraction process and to detect atypical points.

Yahoo! Search and *Exalead* are the only search engines which show the differences between the three types of web mentions (Title mention, URL mention and inlink), since they are the only ones which provide link data. Due to the greater coverage of *Yahoo! Search*, it was selected to give a visual illustration of these differences from a single data source (Figure 1).

Figure 1. Distribution of web mention types according to *Yahoo!*

The Title mention describes an unstable line with high fluctuations (σ=820,490) in relation to inlinks (σ=131,084) due to the fact that an institution can have multiple names and different languages. The results reported by the search engines could thus be misrepresented or out of proportion. It might also be added that Title mentions produce many more results than URL mentions and inlinks, because Title mentions may express different contexts such as authorship, references, acknowledgements, lists, etc., which increases frequency.

On the other hand, the URL mention exhibits fewer variations (σ=246,929) and a similar amount of results as inlinks, but displays several atypical points such as the *Universidad de Murcia* (um.es), the *Universidad de Sevilla* (us.es), the *Universitat Ramón Llull* (url.es) and the *Universitat Internacional de Catalunya* (unica.es). This shows one of the limitations of the URL mention when it comes to be extracted: it includes e-mail addresses (i.e. @ipb.ucm.es), dynamic pages on web traffic services (i.e. http://www.alexa.com/siteinfo/harvard.edu) and other URLs that include the same text as the URL being searched.

This limitation is the case of the four universities mentioned above, which have a domain text similar to, or included in, other web domains. For example, the *Universidad de Murcia* (um.es) is included in multiple web domains (i.e., “orbitum.es”, “botanicum.es”, etc.). The same effect is easily detected for the *Universidad de Sevilla* (us.es) in other domains (i.e., “visit-us.es”, “globalus.es”, etc.). These limitations make it necessary to carry out a prior cleaning-up and data checking process.

4.2. *Correlation*

A correlation matrix was calculated to observe the relationships between the different type of web mentions and the similarities between distinct data sources (table 1). The atypical results of the four previous universities were removed to make the results more accurate.

Table 1. Correlation matrix of the different web citations and web sources (Spearman’s rank correlation coefficient; in bold *ρ*>.9)

The highest correlations are presented in bold. Prior to analysing these data, it is important to mention that the obtained counts from *Yahoo! Search* and *Bing* in 2011 November are exactly the same. This is because they are completely correlated (*ρ*=1) and they present the same correlations with other web sources. Due to this, henceforth we do not mention the correlation between *Yahoo! Search* and *Bing*.

At first glance, Table 1 shows that the highest correlations are between the same types of web mention obtained from different web sources. Hence the highest correlation of Title mentions are between *Yahoo!*/*Bing* and *Exalead* (*ρ*=.945), while the best correlation between Title mentions and URL mentions is obtained from *Google Scholar* (*ρ*=.866) and *Yahoo!* (*ρ*=.840). In the case of inlinks, the Title mentions present high correlations with *Yahoo!* (*ρ*=.816).

As with the Title mentions, the URL mentions correlate better with URL mentions from different search engines than with other web mentions from the same source. The best correlations are between *Yahoo!*/*Bing* and *Google* (*ρ*=.945) and *Exalead* (*ρ*=.921).

According to other web mentions, the URL mentions correlate better with inlinks than with Title mentions. For example, the *Google Scholar* URL mentions present a high correlation with *Yahoo! Search* inlinks (*ρ*=.917), and *Google* with YSE (*ρ*=.945), while the best correlation with Title mentions is between the *Google Scholar* URL mentions and *Exalead* Title mentions (*ρ*=.881).

The great similitude between *Yahoo!* inlinks and *Google Scholar* URL mentions may be due to the fact that *Google Scholar* is an academic search engine that covers papers and patents in which the URL mentions are, in many cases, links.

Finally, regarding inlinks, the very high correlation between *Yahoo!* and YSE (ρ=.990) confirms that both are fed from the same database, with the only difference that *Yahoo! Search* rounds off the data and YSE does not. It is also interesting to mention that *Exalead* inlink counts are the most unstable results because its correlations are rather low (ρ<.51*)* and not comparable with the other inlink sources.

Moreover, the Friedman test for non-parametric ranks was used to confirm the differences between the different web mention types. As the correlation suggests, the URL mentions are closer to the inlinks than to the Title mentions. In the case of *Yahoo!* (the only search engine that allows the three indicators to be obtained), the Friedman test shows that there are significant differences between the three web mention counts (Q=230.214 p-value<.0001).

Since *Exalead* results are very unstable, only *Yahoo!* gives an accurate comparison between all web citation indexes from the same source. In this sense, the *post hoc* analysis of the Nemenyi test is applied to all different indicators retrieved by *Yahoo!* (*Yahoo!* inlink, *Yahoo! Site Explorer*, *Yahoo!* URL mention, and *Yahoo!* Title mention). The results (shown in Table 2) show that inlinks and URL mentions are grouped together whereas Title mentions are segregated from the other two. Additionally, Table 2 provides the same analysis, applied to the indicators recovered by *Google* and *Google Scholar* (URL and Title mentions), showing that URL and Title mentions group together in academic environments (*Scholar*) but not in the general search engine (Table 2).

Table 2. Differences between web citation indexes from a same source (Nemenyi *post hoc* test)

In Table 2 it can be also be observed that the “Mean” for *Yahoo!* Title mention almost duplicates the *Yahoo!* URL mention (this effect is amplified in *Google* URL and Title mentions); the “Mean of Ranks” is higher for Title mentions as well. This difference may be due to the fact that the inlinks and URL mentions are web mentions that point to the source of information, whereas the Title mentions are more ambiguous and are only references to an institution which do not involve a true citation to their web domain. Moreover, an institution can be named in different ways and in different languages with different acronyms, which would produce very different results, especially in a system such as the Spanish one.

4.3. *Regression model*

Due to the lack of search engines and web services that allow information to be obtained about the inlinks that point to a certain website or domain, we postulate whether there is any possibility of estimating the number of inlinks that a Spanish web domain receive from the number of Title and URL mentions, and in what proportion these web mentions allow the number of inlinks to be predicted, and how reliable they are.

To answer these questions, a multiple regression model is applied between the three web mention types from the same data source (only possible with *Yahoo! Search*). The four atypical points were previously removed to reinforce the validity of the results (table 3).

Table 3. Regression analysis model of *Yahoo!* inlinks

The *t*-value coefficient makes it possible to estimate the relationship between *Yahoo!* web indicators (table 3a). The model equation obtained is shown below [eq1]:

$Inlinks = 0.73 + 0.15 \cdot Title + 0.76 \cdot URL$ [eq1];

This model rejects the *Yahoo!* Title mention because its coefficient is not statistically significant (*p*-value=.12). Therefore this variable was removed from the model, thus predicting the number of inlinks only from the number of URL mentions (table 3b). The new model obtained the equations with the following standardised coefficient [eq2] and the unstandardised coefficients [eq3]:

$Inlinks = URL^{.908}$ [eq2];

$Inlinks = 1.42 + 0.86 \cdot URL$ [eq3];

In this case, the *Yahoo!* URL mentions obtain an adjusted R^2=.82. This means that these web mentions may explain and predict in 82% of cases the number of inlinks that a website receives. It is also interesting to note that the coefficient is close to 1 which suggests that the URL mention values are quite similar to the inlinks.

5. Discussion

The correlation analysis shows that the closest web mention alternative to inlinks is URL mention as it may better express the transitivity of a hyperlink. These similar results were described by Thelwall and Sud (2011) who found that URL citation was the measure that correlated best with inlinks, concluding that the different web mention types could be used interchangeably for impact measurements, although there will be differences in their results. Along these lines, Thelwall, Sud and Wilkinson (2012) also observed that URL citation is the best type of data for co-link analysis.

The similarity of these results with the findings obtained in the analysis of the Spanish university system reinforces the hypothesis that these different web mention types could be used as a proxy to measure the web impact and visibility of a website on the Web.

The regression analysis has also confirmed this hypothesis, finding that the URL mention is the unique estimator that explains the inlinks. Its coefficient value (and the data displayed in Figure 1) shows that URL mentions and inlink counts are rather similar; therefore, it could be concluded that, although these mention counts correlate highly with the inlink counts, URL mentions are enough to estimate the incoming links that a website receives.

Furthermore, although the correlation achieved between all the web mention types considered is high, the analysis also found that the different web mentions are better correlated among themselves when they come from different web sources than when they are between the other types of web mentions extracted from the same search engine. For example, the correlation between *Google* URL mentions and *Yahoo!* URL mentions is very high (R=.945), and exactly the same as between *Google* URL mentions and *Yahoo!* inlinks.

Limitations of the Spanish university system

The study of web mentions (both Title and URL mentions) through the Spanish university web domains introduces some important peculiarities that must be considered as they differ slightly from the results obtained in the previous studies carried out by

Thelwall and Sud (2011) and Thelwall, Sud and Wilkinson (2012). The main considerations are shown below:

Title mentions

Title mentions provide unstable results with a high variability caused by the multiple name variants which an institution can present, such as acronyms and other language versions, due to the fact that some Spanish universities can be named in different ways because there are four official languages.

This effect is reflected in the correlation obtained between title mentions and *Yahoo!* inlinks, which, although it is high (with *Google*: R=.62; with *Yahoo!*: R=.82), is lower than that achieved in the aforementioned studies.

In this study, the solution was to combine the mentions of each language variant, although the number of possible combinations is elevated, making this procedure especially difficult to apply in countries with higher language diversity. This means that some specific university systems may be affected to a greater extent than others if this indicator is applied on a global basis.

Furthermore, this indicator is rather ambiguous because it may signify authorship, reference, a list or acknowledgement.

URL mentions

The measurement of URL mentions is more stable and produces results closer to inlink counts, but this study has identified and confirmed two important limitations with respect to URL mentions, also previously detected by Thelwall and Sud (2011).

The first limitation is related to the presence of atypical points due to the fact that some short domain names may be included in other longer URLs (especially in e-mail addresses) when these data are extracted. This limitation is problematic because it is hard to avoid and produces atypical points which distort correlations and can exaggerate the count of a website.

The online academic systems treated in previous studies (USA, UK and China) present particularities in the university URL syntaxes (not shared by the Spanish academic system) that minimise this effect. On one hand, United States uses the ".edu" domain for academic institutions. On the other hand, UK and China have a specific second-level domain for academic institutions environments (".ac.uk" and "edu.cn" respectively). This procedure facilitates the calculation of URL mentions.

The Spanish academic system does not have any specific web domain for universities. As a matter of fact, the multi-domain (the maintenance of different official URLs) is a common practice (Orduña-Malea, in press), which maximises this effect.

In this study, we have been able to remove the atypical points produced by the first limitation. Moreover, for each academic URL the external inlinks from their URL alias have been rejected, and then all aliases belonging to the same university have been aggregated obtaining a unique value for each institution. This procedure (which is not necessary in other academic systems) has reinforced the results and allowed the design of a consistent regression model that has determined the relationship between the new proposed indicators and inlinks.

The second limitation is derived from the previous one, and is related to the fact that some of the URL mentions -assuming that they come from the required institution and cannot be considered as noise- do not express the same meaning as inlinks. For example, e-mail addresses cannot be understood as a visibility or transitivity measurement. This constitutes a problem inherent in data extraction affecting every academic web domain, so it does not influence the correlations.

It has not been possible to control this second limitation and, therefore, some of these claims have to be cautious because they may affect the meaning and interpretation of the URL mention measurements.

6. Conclusions

The main conclusions are set out below:

a) The different web mentions are better correlated between themselves when they come from different web sources than when they are between the other types of web mentions extracted from the same search engine.

b) The web mentions (both title and URL) achieve higher correlations with *Yahoo!* inlinks in the Spanish academic system. Notwithstanding, in order to predict inlinks, URL mentions are enough to predict (in 82% of cases) the number of inlinks that a website receives whereas the title mentions should be rejected.

c) Despite the higher correlations obtained, both title and URL mentions exhibit certain limitations:
   - Title mentions depend on language diversity, so different academic systems may be affected differently.
   - URL mentions can present a great amount of noise (the URL text may be embedded in other URLs outside the institution under analysis), generating atypical points. Additionally, these URL mentions may express concepts different from those expressed by inlinks, so that their interpretation should be treated with caution.

d) The study demonstrates that the previously expressed limitations increase in the context of the Spanish academic web system:
   - The existence of different official names for some Spanish universities makes title mention indicators unsuitable for predicting inlinks.
   - Multi-domain practice maximises the noise in URL mention results, which need advanced search queries and cleaning-up processes.

Considering the results obtained, it may be concluded that URL mentions are the best indicators to substitute inlinks, but we also advise caution with anomalous results. Even so, although this indicator can be interpreted as a measure of web visibility, it does not necessarily mean a link relationship or a navigational reference from a website, since its interpretation is not exactly the same as an inlink.

Therefore, as a general conclusion, it can be stated that the web mention indicators correlate highly with inlinks, but their limitations make them prone to environmental influence to a great extent (language, web domain policy, etc.). This effect (and the time-consuming steps to avoid it) makes these indicators (as they currently exist) inadequate for use on a global basis (such as a World University Ranking).

Further research

Due to the aforementioned problems, further research is necessary to avoid these limitations, preferably in an automatic manner. The manual cleaning-up of data makes this procedure useless in the analysis of a wide range of universities.

Likewise, the comprehensive analysis of other problematic academic systems should help to establish a clearer relationship between these web indicators. Furthermore, determining the percentage of noise in URL mention results may provide further insights into the prediction of inlinks.

Finally, the appearance of new search engines which provide advanced inlink commands, and the expansion of functionalities of current sources such as *Ahrefs*, *Open Site Explorer* or *Majestic SEO* should be followed and comprehensive analysed.

7. Endnotes

[1] Ranking Web of World Universities. Available at:
http://www.webometrics.info (accessed 10 January, 2013).
[2] 4 International Colleges & Universities. Available at:
http://www.4icu.org (accessed 10 January, 2013).
[3] Open Site Explorer. Available at:
http://www.opensiteexplorer.org (accessed 10 January, 2013).
[4] Majestic SEO. Available at:
http://www.majesticseo.com (accessed 10 January, 2013).
[5] Ahrefs. Available at:
http://ahrefs.com (accessed 10 January, 2013).
[6] WISER Project. Available at:
http://www.wiserweb.org (accessed 10 January, 2013).
[7] Web Indicators Portal. Available at:
http://www.webindicators.org (accessed 10 January, 2013).
[8] EICSTES Project. Available at:
http://www.eicstes.org (accessed 10 January, 2013).
[9] ACUMEN Project. Available at:
http://research-acumen.eu (accessed 10 January, 2013).

8. References

Adecannby, J. (2011), “Web link analysis of interrelationship between top ten African universities and world universities”, *Annals of library and information studies*, Vol. 58, pp. 128-138.

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. 234–244.

Aguillo, I., 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*, Vol. 57, No.10, pp. 1296-1302.

Aguillo, Isidro F. (2009), “Measuring the institutions’ footprint in the web”, *Library Hi Tech*, Vol. 27, No.4, pp. 540–556.

Barabasi, A. L. and Albert, R. (1999), “Emergence of Scaling in Random Networks”, *Science*, Vol. 286, No. 5439, pp. 509-512.

Bar-Ilan, J. (2003), “The use of web search engines in information science research”, *Annual review of information science and technology*, Vol. 38, pp. 231-288.

Bar-Ilan, J. (2004), “A microscopic link analysis of academic institutions within a country – The case of Israel”, *Scientometrics*, Vol. 59, No. 3, pp. 391–403.

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.

Björneborn, L., and Ingwersen, P. (2004), “Toward a basic framework for webometrics”, *Journal of the American Society for Information Science and Technology*, Vol. 55, No.14, pp. 1216–1227.

Bland, J. M. and Altman, D.G. (1996), “Transforming data”, *British Medical Journal,* Vol. 312, pp. 770.

Boudourides, M. A., Sigrist, B. and Alevizos, P. D. (1999). Webometrics and the self-organization of the European information society. *Rome Meeting of the SOEIS Project.*

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A. and Callahan, E. (1998), “Invoked on the Web”, *Journal of the American Society for Information Science,* Vol. 49, No. 14, pp.1319–1328.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.

Harries, G., Wilkinson, D., Price, E., Fairclough, R. and Thelwall, M. (2004), “Hyperlinks as a data source for science mapping”, *Journal of Information Science,* Vol. 30, No. 5, pp. 436–447.

Heimeriks, G. and Van den Besselaar, P. (2006), “Analyzing hyperlink networks: The meaning of hyperlink based indicators of knowledge”, *Cybermetrics*, Vol. 10.

Kousha, K. and Horri, A. (2004), “The relationship between scholarly publishing and the counts of academic inlinks to Iranian university Web sites: Exploring academic link creation motivations”, *Journal of Information Management and Scientometrics*, Vol. 1, No. 2, pp. 13-22.

Kousha, K. and Thelwall. M. (2009), “Google Book Search: Citation Analysis for Social Science and the Humanities”, *Journal of the American Society of Information Science and Technology,* Vol. 60, No. 8, pp.1537-1549.

Kretschmer, H. and Aguillo, I. F. (2004), “Visibility of collaboration on the Web”, *Scientometrics*, Vol. 61, No. 3, pp. 405-426.

Nemenyi, P.B. (1963), *Distribution-free Multiple Comparisons,* Princeton University, NJ, USA.

Orduña-Malea, E. (2012), “Fuentes de enlaces web para análisis cibermétricos (2012)”, *Anuario Thinkepi*, Vol. 6, pp. 276-280.

Orduña-Malea, E. (in press), “Espacio universitario español en la Web (2010): estudio descriptivo de instituciones y productos académicos a través del análisis de subdominios y subdirectorios”, *Revista española de documentación científica*.

Orduña-Malea, E. and Ontalba-Ruipérez, J-A. (2012), “Proposal for a multilevel university cybermetric analysis model”, *Scientometrics*.

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, José L. and Aguillo, Isidro F. (2008), “Visualization of the Nordic academic web: Link analysis using social network tools”, *Information Processing & Management*, Vol. 44, No. 4, pp. 1624-1633.

Ortega, José L. and Aguillo, Isidro F. (2009a), “Análisis estructural de la web académica iberoamericana”, *Revista española de documentación científica*, Vol. 32, No. 3, pp. 51-65.

Ortega, José L. and Aguillo, Isidro F. (2009b), “North America academic web space: Multicultural Canada vs. The United States homogeneity”. *ASIST & ISSI Pre-Conference Symposium on Informetrics and Scientometrics*.

Ortega, José L., Aguillo, Isidro F., Cothey, V. and Scharnhorst, A. (2008), “Maps of the academic Web in the European Higher Education Area: an exploration of visual Web indicators”, *Scientometrics*, Vol. 74, No. 2, pp. 295-308.

Qiu, J., Cheng, J. and Wang, Z. (2004), “An analysis of backlinks counts and web impact factors for Chinese university websites”, *Scientometrics*, Vol. 60, No. 3, pp. 463-473.

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.

Seidman, E. (2007), “We are flattered, but…”, *Bing Community*. Available at: http://www.bing.com/community/site_blogs/b/search/archive/2007/03/28/we-are-flattered-but.aspx (accessed 20 October, 2012).

Smith, A. and Thelwall, M. (2002), “Web Impact Factors for Australasian Universities”, *Scientometrics*, Vol. 54, No. 3, pp. 363-380.

Smith, A. G. (1999), “A tale of two web spaces: comparing sites using web impact factors”, *Journal of documentation*, Vol. 55, No. 5, pp.577–592.

Stuart, D. and Thelwall, M. (2006), “Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry”, *Research Evaluation*, Vol. 15, No. 2, pp. 97-106.

The Washington Post (2009), “It's Official: Yahoo-Microsoft Announce 10-Year Search/Ad Pact”, available at: http://www.washingtonpost.com/wp-dyn/content/article/2009/07/29/AR2009072901108.html (accessed 27 February 2013).

Thelwall, M. (2001), “Extracting macroscopic information from web links”, *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 13, pp.1157–1168.

Thelwall, M. (2002), “An initial exploration of the link relationship between UK university Web sites”, *ASLIB Proceedings*, Vol. 54, No. 2, pp. 118-126.

Thelwall, M. and Aguillo, Isidro F. (2003), “La salud de las web universitarias españolas”, *Revista española de documentación científica*, Vol. 26, No.3, pp. 291-305.

Thelwall, M. and Sud, P. (2011), “A comparison of methods for collecting web citation data for academic organisations”. *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 8, pp. 1488-1497.

Thelwall, M. and Sud, P. (2012), “Webometric research with the Bing search API 2.0”. *Journal of informetrics*, Vol. 6, No. 1, pp. 44-52.

Thelwall, M. and Zuccala, A. (2008), “A university-centred European Union link analysis”, *Scientometrics*, Vol. 75, No. 3, pp. 407–420.

Thelwall, M., Binns, R., Harries, *G.*, Page-Kennedy, T., Price, E. and Wilkinson, D. (2002), “European Union associated university Websites”, *Scientometrics*, Vol. 53, No. 1, pp. 95-111.

Thelwall, M., Sud, P. and Wilkinson, D. (2012), “Link and co-inlink network diagrams with URL citations or title mentions”, *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 10, pp. 1960-1972.

Thelwall, M., Tang, R., and Price, E. (2003), “Linguistic patterns of academic web use in Western Europe”, *Scientometrics*, Vol. 56, No. 3, pp. 417–432.

Thelwall. M. and Kousha, K. (2008), “Online Presentations as a Source of Scientific Impact?: An Analysis of PowerPoint Files Citing Academic Journals”, *Journal of the American Society of Information Science and Technology*, Vol. 59, No. 5, pp. 805-815.

Vaughan, L. (2012), “An alternative data source for web hyperlink analysis: ‘sites linking in’ at Alexa Internet”, *Collnet journal of scientometrics and information management*, Vol. 6, No. 1.

Vaughan, L. and Romero-Frias, E. (2012)**, "**Exploring Web keyword analysis as an alternative to link analysis: a multi-industry case", *Scientometrics*, Vol. 93, No. 1, pp. 217-232.

Vaughan, L. and Shaw, D. (2003), "Bibliographic and Web citations: What is the difference?", *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 14, pp. 1313-1322.

Vaughan, L. and Yang, R. (2012), "Web data as academic and business quality estimates: A comparison of three data sources", *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 10, pp. 1960-1972.

Vaughan, L. and You, J. (2010), "Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept", *Journal of Informetrics,* Vol. 4, No. 4, pp. 483-491.

Vaughan, L., Kipp, M. and Gao, Y. (2007), "Why are Websites co-linked? The case of Canadian Universities", *Scientometrics*, Vol. 72, No. 1, pp. 81–92.

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.

Zhang, Y. (2006), "The Effect of Open Access on Citation Impact: A Comparison Study Based on Web Citation Analysis", *Libri*, Vol. 56, No. 3, pp.145-156.

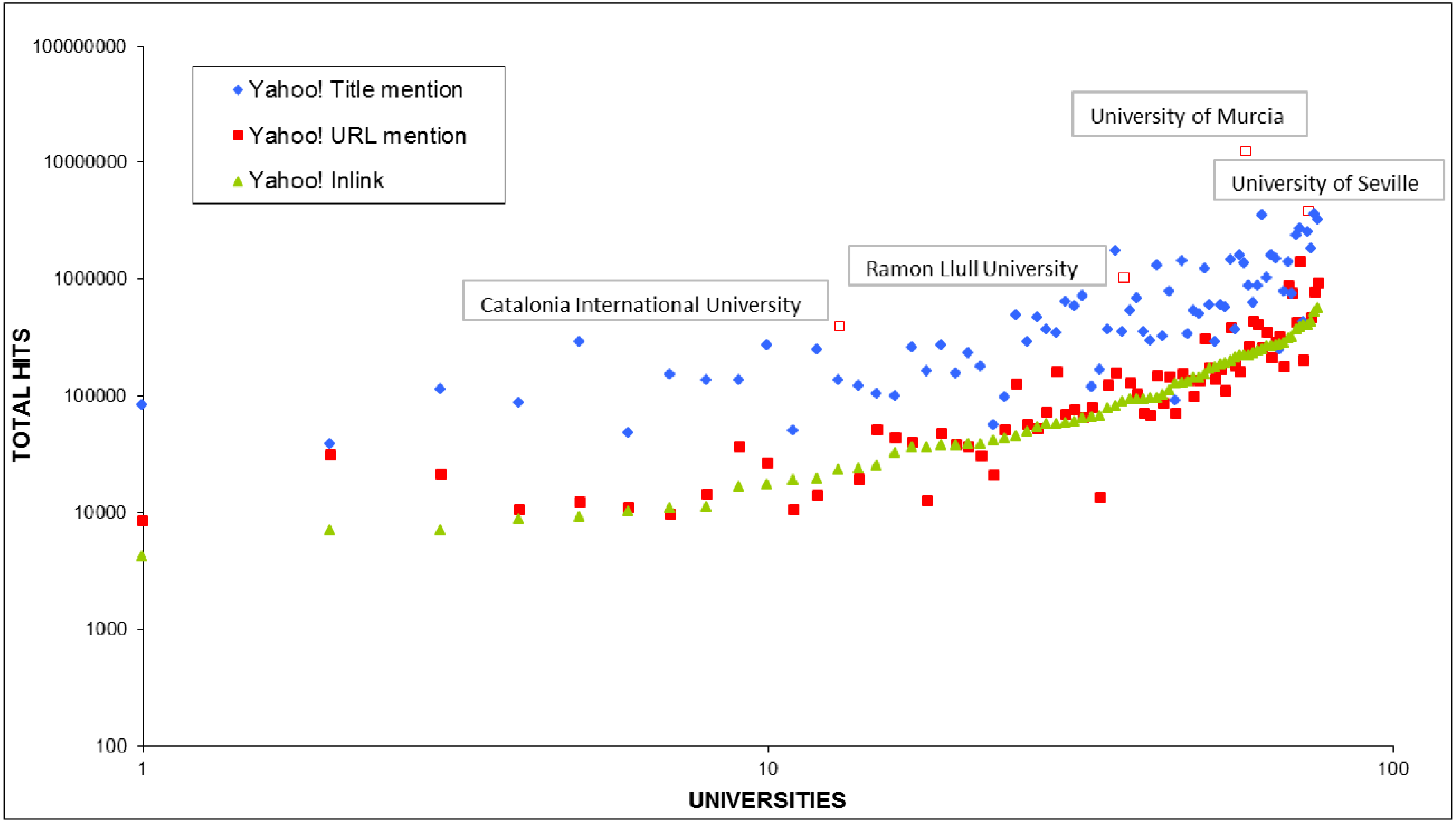

Figure 1. Distribution of web mention types according to *Yahoo!*

Table 1. Correlation matrix of the different web citations and web sources (Spearman' rank correlation coefficient; in bold *ρ*>.9)

Variables	Google Title mention	Bing Title mention	Yahoo! Title mention	Exalead Title mention	Scholar Title mention	Google URL mention	Bing URL mention	Yahoo! URL mention	Exalead URL mention	Scholar URL mention	Yahoo! Inlink	Exalead Inlink	YSE Inlink
Google Title mention	1												
Bing Title mention	.684	1											
Yahoo! Title mention	.684	**1.000**	1										
Exalead Title mention	.737	**.945**	**.945**	1									
Scholar Title mention	.700	.852	.852	.856	1								
Google URL mention	.624	.806	.806	.821	.826	1							
Bing URL mention	.593	.840	.840	.835	.792	**.945**	1						
Yahoo! URL mention	.593	.840	.840	.835	.792	**.945**	**1.000**	1					
Exalead URL mention	.616	.771	.770	.813	.805	**.922**	**.921**	**.921**	1				
Scholar URL mention	.669	.868	.867	.881	.866	**.931**	**.939**	**.939**	**.921**	1			
Yahoo! Inlink	.623	.816	.816	.827	.787	**.945**	**.933**	**.933**	**.925**	**.917**	1		
Exalead Inlink	.311	.387	.387	.388	.342	.422	.510	.510	.510	.447	.410	1	
YSE Inlink	.636	.807	.807	.827	.794	**.945**	**.934**	**.934**	**.925**	**.916**	**.990**	.403	1

Table 2. Differences between web citation indexes from a same source (Nemenyi *post hoc* test)

Yahoo! Search

Sample	Frequency	Mean	Mean of ranks	Groups
Yahoo! inlink	76	138,789.3	1.849	A
YSE inlink	76	145,734.6	1.921	A
Yahoo! URL mention	76	407,983.1	2.362	A
Yahoo! title mention	76	734,588.6	3.868	B

Google* and *Google Scholar

Sample	Frecuency	Mean	Mean of ranks	Groups
Scholar URL	76	10,908.3	1,382	A
Scholar Mention	76	16,389.9	1,618	A
Google URL	76	1,611,346.9	3,066	B
Google Mention	76	7,141,153.9	3,934	C

Table 3. Regression analysis model of Yahoo! inlinks

a) Multiple regression analysis model of Yahoo! inlinks according to web mentions

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. error	Beta		
(Constant)	.734	.694		1.057	.294
Yahoo! URL mention	.756	.083	.795	9.130	.000
Yahoo! Title mention	.148	.094	.137	1.574	.120

Adjusted R^2=.83

b) Simple regression analysis model of Yahoo! inlinks according only to URL mentions

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. error	Beta		
(Constant)	1.420	.546		2.599	.011
Yahoo! URL mention	.864	.048	.908	18.123	.000

Adjusted R^2=.82