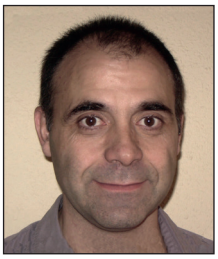
RECEIVED CITATIONS AS A MAIN SEO FACTOR OF *GOOGLE SCHOLAR* RESULTS RANKING

Las citas recibidas como principal factor de posicionamiento SEO en la ordenación de resultados de *Google Scholar*

Cristòfol Rovira, Frederic Guerrero-Solé and Lluís Codina

Nota: Este artículo se puede leer en español en:
http://www.elprofesionaldelainformacion.com/contenidos/2018/may/09_esp.pdf

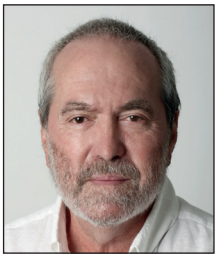


Cristòfol Rovira, associate professor at *Pompeu Fabra University* (*UPF*), teaches in the *Departments of Journalism and Advertising*. He is director of the master's degree in *Digital Documentation* (*UPF*) and the master's degree in *Search Engines* (*UPF*). He has a degree in Educational Sciences, as well as in Library and Information Science. He is an engineer in Computer Science and has a master's degree in Free Software. He is conducting research in web positioning (SEO), usability, search engine marketing and conceptual maps with eyetracking techniques.
https://orcid.org/0000-0002-6463-3216

cristofol.rovira@upf.edu

Frederic Guerrero-Solé has a bachelor's in Physics from the *University of Barcelona* (*UB*) and a PhD in Public Communication obtained at *Universitat Pompeu Fabra* (*UPF*). He has been teaching at the *Faculty of Communication* at the *UPF* since 2008, where he is a lecturer in Sociology of Communication. He is a member of the research group *Audiovisual Communication Research Unit* (*Unica*).
https://orcid.org/0000-0001-8145-8707

frederic.guerrero@upf.edu



Lluís Codina is an associate professor in the *Department of Communication* at the *School of Communication*, *Universitat Pompeu Fabra* (*UPF*), Barcelona, Spain, where he has taught information science courses in the areas of Journalism and Media Studies for more than 25 years. He is the director of the master's in Social Communication Program at *UPF*. He holds a PhD in Journalism from the *Autonomous University of Barcelona*, Spain, where he was an assistant professor. He is lecturing in the master's in *Online Documentation and Search Engine Optimization* at the *UPF-Barcelona School of Management*, Spain.
https://orcid.org/0000-0001-7020-1631

lluis.codina@upf.edu

Universitat Pompeu Fabra, Departament de Comunicació
Roc Boronat, 138. 08018 Barcelona, Spain

Summary

The aim of this article is to analyze the web positioning factors that can influence the order, by relevance, in *Google Scholar* and the subsequent evaluation of the importance of received citations in this ordering process. The methodology of reverse engineering was applied, in which a comparison was made between the *Google Scholar* ranking and another ranking consisting of only the number of citations received by documents. This investigation was conducted employing four types of searches without the use of keywords: by publication, year, author, and "cited by". The results were matched in the four samples with correlation coefficients between the two highest rankings, which exceeded 0.9. The present study demonstrates more clearly than in previous research how citations are the most relevant off-page feature in the ranking of search results on *Google Scholar*. The other features have minimal influence. This information provides a solid basis for the academic search engine optimization (ASEO) discipline. We also developed a new analysis procedure for isolating off-page features that might be of practical use in forthcoming investigations.

Manuscript received on 24-11-2017
Accepted on 13-04-2018

Keywords

ASEO; SEO; Reverse engineering; Citations; *Google Scholar*; Indicators; Rankings; Algorithms; Academic search engines.

Resumen

El objetivo de este artículo es analizar los factores de posicionamiento (SEO) externos que pueden influir en la ordenación por relevancia en *Google Scholar* y luego identificar el peso las citas recibidas en esta ordenación. Se ha aplicado una metodología de ingeniería inversa comparando el ranking de *Google Scholar* con un ranking formado tan sólo por el número de citas recibidas por los documentos. El estudio se realizó a partir de cuatro tipos de búsquedas sin palabras clave: por publicación, año, autor y "citado por". Los resultados fueron coincidentes en las cuatro muestras con coeficientes de correlación entre los dos rankings superiores al 0,9. El presente estudio demuestra de forma más clara que en anteriores investigaciones que las citas recibidas es el factor SEO externo más relevante en el ranking de los resultados en *Google Scholar*. Los demás factores tienen una influencia mínima. Esta información proporciona una base sólida para la disciplina del posicionamiento en buscadores académicos (ASEO). También hemos desarrollado una nueva propuesta metodológica que aísla los factores SEO externos y que puede ser útil en futuras investigaciones.

Palabras clave

ASEO; SEO; Ingeniería inversa; Citaciones; *Google Scholar*; Indicadores; Rankings; Algoritmos; Ordenación; Motores de búsqueda académicos.

Rovira, Cristòfol; **Guerrero-Solé, Frederic**; **Codina, Lluís** (2018). "Received citations as a main SEO factor of *Google Scholar* results ranking". *El profesional de la información*, v. 27, n. 3, pp. 559-569.

https://doi.org/10.3145/epi.2018.may.09

1. Introduction

Search engine optimisation (SEO) is the process employed to optimise websites and their content to place them in favourable positions in search engine results (**Enge**; **Spencer**; **Stricchiola**, 2015). SEO is also a well-established profession within the new industry of digital communication, as shown by the existence of a wide range of monographs, professional publications and academic work. Its purpose is to highlight and strengthen the quality of documents to increase their visibility to the algorithms that establish the ranking positions in search engines, particularly *Google*. This goal must be achieved without falsifying the characteristics of documents, i.e., without employing fraudulent means.

Google Search results pages are ordered by relevance (*Google*, 2017). According to *Google*, this relevance criterion is calculated based on more than 200 features. *Google* does not specify these features or their specific weight; they merely disclose partial and general information, including that the quality of the content and backlinks are the two predominant factors (**Ratcliff**, 2016; **Schwartz**, 2016).

> The reason provided by *Google* for this lack of transparency is to fight against spam

The reason provided by *Google* for this lack of transparency is to fight against spam (**Beel**; **Gipp**, 2010). If all of the details of ranking factors were made available, spammers could more easily place low-quality documents in favourable positions. Nevertheless, this black box policy works to the detriment of SEO professionals who conduct their activities ethically and whose work is hindered by a lack of reliable information.

Some SEO companies (**Gielen**; **Rosen**, 2016; *Localseoguide*, 2016; *MOZ*, 2015; *Searchmetrics*, 2016) conduct reverse engineering research to measure the impact of the factors involved in *Google*'s positioning process. In this research, many searches have been analysed to identify positioning factors based on the characteristics of pages placed in the first positions. Due to the great number of factors involved in the process of positioning, it is extremely difficult to establish the factors that are truly relevant and the extent to which they influence the final positioning of documents. In addition, *Google*'s positioning process is highly dynamic, with the algorithm undergoing dozens of changes per year (*MOZ*, 2017).

> The ASEO (*academic search engine optimisation*) is the SEO applied to academic search engines

In recent years, SEO has been applied to academic search engines. This new process is known as Academic SEO (ASEO) (**Beel**; **Gipp**, 2009b, 2010; **Codina**, 2016; **Martín-Martín** *et al.*, 2016a; **Muñoz-Martín**, 2015). Scholars are placing increasingly greater emphasis on enhancing the visibility of their articles in academic search engines. Articles appearing in the leading positions enhance their visibility, thus increasing the probability of being read and cited, and as a consequence, they are more likely to improve the personal h indices of their authors (**Farhadi** *et al.*, 2013).

In many cases, the same optimisation procedures used successfully on *Google Search* are being applied to *Google Scholar*. However, *Google Scholar* has its own algorithm. Few studies have addressed the specific ordering factors employed by *Google Scholar*, and among those that could

be cited are **Beel** and **Gipp** (2009b; 2009c; 2010); **Beel**, **Gipp** and **Wilde** (2010); **Martín-Martín** *et al.* (2014; 2017); **Orduña-Malea** *et al.* (2016).

The purpose of the present study was to analyse the features of the documents that can influence relevance rankings in *Google Scholar*. We are particularly interested in the citations received by documents. We aimed to assess the influence of the number of citations received in the ranking algorithm. The number of times that a document is cited is a key feature for determining the specificity of the *Google Scholar* ranking process. We believe that the influence of citations is much greater than authors and publishers might believe. For example, the instructions to authors from academic journals provide guidelines regarding how to improve their ranking positions in *Google Scholar* (*Elsevier*, 2012; *Wiley*, 2015; *Emerald Publishing Limited*, 2017). In these guides, the citations received are not mentioned or are treated without the importance that they deserve.

This article reports the findings of a reverse engineering study that used a new method of analysis. This method allows us to block some factors of the algorithm of positioning, specifically those depending on external elements of ranked pages. In this manner, we could focus the study on a small set of factors with greater control.

Our hypothesis is that if we compare the rankings applying only the number of citations received with the standard *Google Scholar* ranking in searches in which only external factors participate, then we can identify the weight of the citations in the set of these external factors. If the two compared rankings are similar, then the citations will carry significant weight.

This new methodology is possible because of *Google Scholar*'s advanced search form, which allows users to restrict the search fields to the author, year and source. Only external factors participate in these types of searches in which there are no keywords. In this way, the results obtained herein are far more reliable than those of previous studies using reverse engineering on *Google Search* without this control of variables.

2. Related works

Google Scholar has become an alternative to classic scientific citation indexing services, such as *Web of Science* (*WoS*) or *Scopus*. The positions of these commercial indexing services in the market could be jeopardised if *Google Scholar* offers a free product of similar quality. For this reason, *Google Scholar* has been analysed using several approaches:

- Comparative linear or coverage analysis, aiming to establish its quality and utility (**Giustini**; **Boulos**, 2013; **Walters**, 2008; **De-Winter**; **Zadpoor**; **Dodou**, 2014; **Harzing**, 2013; 2014; **De-Groote**; **Raszewski**, 2012; **Orduña-Malea** *et al.*, 2014; 2015; **Pedersen**; **Arendt**, 2014; **Jamali**; **Nabavi**, 2015);
- Assessment of the impact of the authors, their citations or H indices (**Van-Aalst**, 2010; **Jacsó**, 2008a; 2008b; 2009; 2012; **Martín-Martín** *et al.*, 2014; 2017; **Farhadi** *et al.*, 2013); and
- Assessment of the utility of *Google Scholar* for bibliometric studies regarding the quality of the scientific activity (**Aguillo**, 2012; **Jacsó**, 2009; **Torres-Salinas**; **Ruiz-Pérez**; **Delgado-López-Cózar**, 2009; **Beel**; **Gipp**, 2010; **Delgado-López-Cózar** *et al.*, 2012; 2014; **Martín-Martín** *et al.*, 2016b).

Limited research regarding the process of information retrieval and search effectiveness has, however, been conducted (**Jamali**; **Asadi**, 2010; **Walters**, 2008). Few works about the intervening factors in ranking algorithms according to relevance have been published (**Beel**; **Gipp**, 2009a; 2009b; 2009c; **Beel**; **Gipp**; **Wilde**, 2010). Unlike the process of positioning in *Google Search*, that used in *Google Scholar* has aroused little scientific interest, which is somewhat unexpected considering that it influences the articles that are read. It is widely acknowledged that the first items appearing on a search result list receive more attention from users than subsequent items do (**Marcos**; **González-Caro**, 2010). A better position in the ranking implies better chances of being found and read.

> Unlike the process of positioning in *Google Search*, that used in *Google Scholar* has aroused little scientific interest

Some conclusions can be drawn from the existing works regarding relevance rankings in *Google Scholar*:

- The keywords used in the search must appear in the document's title to enable favourable positioning of the document (**Beel**; **Gipp**, 2009a);
- The frequency of keywords in the text of the document does not appear to be a determining factor in establishing its ranking order (**Beel**; **Gipp**, 2009a);
- Recent articles are more highly ranked than older articles (**Beel**; **Gipp**, 2009a) to compensate for the Matthew effect (**Merton**, 1968): articles with many citations tend to be ranked first; therefore, these articles have more readers and more citations and consequently consolidate their positions at the top **Martín-Martín** *et al.*, 2016b); and
- The number of citations received is a determining factor in establishing the ranking order by relevance (**Beel**; **Gipp**, 2009c; **Martín-Martín** *et al.*, 2014).

The latter conclusion is particularly relevant to the present study. However, these investigations have some limitations. In **Beel** and **Gipp** (2009c), all SEO features were analysed together; therefore, the variables related to on-page features were not blocked, and the results are not sufficiently clear. In **Martín-Martín** *et al.* (2014) only searches by year were used.

The central aim of the present research was to corroborate this conclusion by applying a methodology that establishes stricter control over variables. This methodology allowed us to obtain an accurate insight into the relevance of received citations in relation to all external features of the ranking algorithm in *Google Scholar*.

3. Reverse engineering

Reverse engineering is a method of analysis used to obtain information about how a system or device is designed. It is generally used to study electronic devices to identify their components and functioning processes. It is also used to develop software to obtain the font codes from compiled programmes.

Reverse engineering has been applied to *Google* searches to ascertain intervening factors in relevance rankings (*Localseoguide*, 2016; *MOZ*, 2015; *Searchmetrics*, 2016). Using partial information provided by *Google* (2017), search engine results pages have been studied to identify the functioning of ranking algorithms. The characteristics of the documents in the top positions are examined to determine the intervening factors and their weights. Reverse engineering requires great effort since algorithms are complex and subject to constant metamorphosis (**Van-der-Graaf**, 2012).

> The reverse engineering of search engines involves the calculation of correlation coefficients between the position of a page and the values of the factors that supposedly intervene in the algorithm

The process of reverse engineering of search engines generally involves the calculation of Spearman's correlation coefficients between the position of a page in a search and the values of the factors that supposedly intervene in the ranking algorithm. A higher correlation indicates that greater weight can be attributed to the feature undergoing analysis in its contribution to the ranking. A correlation coefficient of 0.4 to 0.7 is generally considered moderate, whereas one greater than 0.7 is considered high. However, in more complex cases, such as the current case, which involves a large number of variables, correlation coefficients rarely exceed 0.3 (*MOZ*, 2015).

Although *Google* does not provide detailed information about how the algorithm's ranking works, it does provide some general information about the features involved. From this partial information, it can be deduced that more than 200 factors are involved in its ordering criteria of relevance. These factors can be divided into two main types: on page (internal) and off page (external):

- The on-page features are related to the content of documents and the presence of the keywords used in searches in the text of these documents (**Enge**; **Spencer**; **Stricchiola**, 2015).
- The off-page features are quality indicators that are related not to the content of the document but rather to its context. Examples of off-page features are the quality and quantity of backlinks counted by means of PageRank (**Maciá-Domene**, 2015).

Previous applications of reverse engineering to the ranking criteria used by *Google Search* have encountered great difficulty in analysing these 200 variables (*Localseoguide*, 2016; *MOZ*, 2015; *Searchmetrics*, 2016).

The situation is even worse in the case of *Google Scholar* since the information provided about its algorithm ranking is even scarcer than that for *Google Search*. One of the few explicit explanations about *Google Scholar*'s algorithm is that:

> "*Google Scholar* aims to rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by as well as how often and how recently it has been cited in other scholarly literature" (*Google*, 2011).

Based on the vagueness of the data available to scholars, it can be affirmed that the algorithm of *Google Scholar* is simpler, and the number of intervening factors is fewer than those for *Google Search* (**Mayr**; **Walter**, 2007; **Torres-Salinas**; **Ruiz-Pérez**; **Delgado-López-Cózar**, 2009). In *Google Scholar* there is no evidence of the presence of many of the factors that intervene in *Google Search* (Table 1).

Fortunately, *Google Scholar* has several advanced search features that are not available on *Google Search*, which allow researchers to control the on-page features. The on-page features are disabled when searches are performed by author, publisher, or year or by means of the "cited by" link. In these cases, the ranking by relevance is established only by applying off-page features. The variables related to on-page

Table 1. SEO features of *Google Search* and *Google Scholar*

Type	SEO factor	*Google Search*	*Google Scholar*
On-page factors	Content relevance: keywords in title, URL, h1, first 100 words	Yes	Yes
	Technical factors: responsive design, loading speed, usability, metadata and structured data, https...	Yes	?
Off-page factors	Backlinks, *PageRank*	Yes	?
	Received citations	Not	Yes
	Author reputacion	Yes	Yes
	Reputation of the publication or domain	Yes	Yes
	Signals from social networks	?	?
	Traffic, CTR	Yes	?
On-page + Off-page	RankBrain	Yes	?
	Machine-learning, artificial intelligence	Yes	?

features do not play a part, as keywords are not used in the search. They are binary searches; thus, a document might or might not be an author, year or published work.

A second factor makes reverse engineering of *Google Scholar* particularly productive. *Google Scholar* itself provides information about the exact values of the citations received by every document, which is one of the main off-page features. This information is very valuable for reverse engineering. If we compare the standard ranking of *Google Scholar* with the order that results from applying only the number of times that a document is cited, we obtain an accurate approximation of the weight of citations in the ranking algorithms. If both rankings are similar, it indicates that citations are an important factor.

> The field search of *Google Scholar* allows us to study the ranking algorithm when only off-page features participate

If we control the on-page features, we can obtain further conclusive evidence. The statistical method that enables us to analyse these data consists of calculating the correlation coefficient between the conventional ranking of *Google Scholar* and the order obtained by merely applying the number of citations.

4. Methodology

In this research, we compared the *Google Scholar* ranking with another ranking created by applying only the number of citations received by the documents in the lists result. We therefore compared a complex ranking that, according to *Google* (2011), considers at least the complete text, the publication, the author and the citations received, with another more basic ranking consisting of only one of these variables, i.e., the citations received. This comparison provides an approximation of the weight of citations in the *Google Scholar* ranking algorithm. If the rankings are very similar, then the number of citations is an important factor.

Nevertheless, the number of citations is an off-page feature. Therefore, the comparison should be made with *Google Scholar* rankings based only on off-page features, i.e., rankings in which on-page features do not participate directly, related to the matching of the keywords in the search and in the document.

To achieve this control of on-page features, we used four types of basic searches in which thematic keywords did not participate. To this end, we used

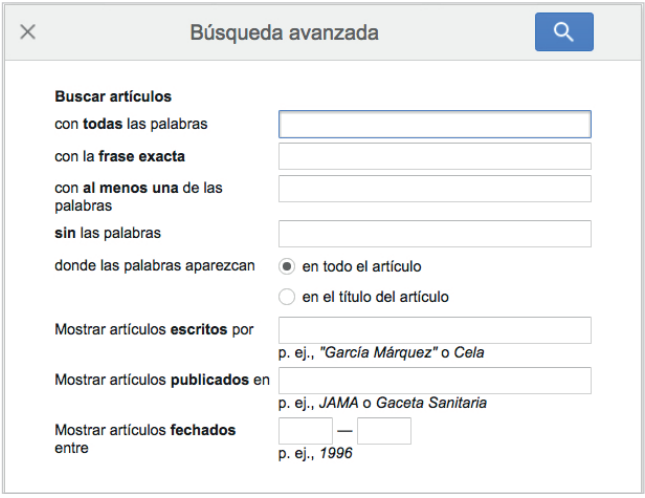


Figure 1. *Google Scholar* advanced search form

the fields for publisher, author and year in *Google Scholar*'s advanced search fields (Figure 1). We also employed a fourth type of non-thematic search using the link "cited by", which is available for every item in the lists of *Google Scholar* results (Figure 2). This link enabled us to conduct a new search and to obtain the works that cite the initial document containing this link.

These data were obtained using the program *Publish or Perish* (**Harzing**, 2011) between 01/10/2016 and 20/02/2017. *Publish or Perish* is a software that automatically extracts the results of searches provided by *Google Scholar*. It is one of the few tools to do so because *Google* does not allow massive extraction of data.

In each of these four search types, we performed 25 searches of 1000 results, thus reaching a total of 100,000 items of data. In each case, we compared the ranking of *Google*

Figure 2. Link "Cited by" in list results of *Google Scholar*

Scholar with the order obtained when considering only the number of citations. For statistical analysis, we applied Spearman's correlation coefficient because the distributions were not normal.

> In searches without keywords in *Google Scholar*, the order by the number of citations received was almost identical to the order by relevance

The samples were chosen to avoid biases. The searches excluded the documents of patents because these documents do not follow the same citation patterns as academic articles. We also endeavoured to ensure that the volume of citations in all of the retrieved documents was relatively high to avoid results with few citations received. The ranking of these uncited results must be ranked according to other factors; therefore, the data would be contaminated. The selection procedures of the four chosen samples were as follows:

- Publications were chosen at random.
- Years. The 25-year period between 1989 and 2013 was chosen. The four years between 2014 and 2017 were excluded to achieve a similar overall volume of citations for all years.
- Authors. The search was performed by surname. The most common surnames in the United Kingdom and United States were selected because they are very common authors in *Google Scholar*. Each search had several authors with the same surname.
- Cited by. The choice of articles in our "cited by" search was performed at random.

Table 2. Global correlation coefficients for each type of search (α = 0.0025)

Type of search	Spearman rho	L	U
Publication	0.999	0.999	0.999
Year	0.909	0.898	0.919
Author	0.998	0.997	0.998
Cited by	0.999	0.999	0.999
Global	0.999	0.999	0.999

The data provided by *Google Scholar* regarding the number of citations of each document from the list of results were transformed into ranges (ordinal scale) (**Beel**; **Gipp**, 2009c). Thus, we created an alternative ranking for each search based only on the number of citations, which was later compared with the standard relevance ranking of *Google Scholar* by means of Spearman's correlation.

To add the data for each type of search and to obtain overall correlation coefficients, we compared the position in *Google Scholar* of each document with the average of the ranking of citations of the 25 samples. We also applied the average to calculate the overall value for the four samples, involving 100,000 analysed data points.

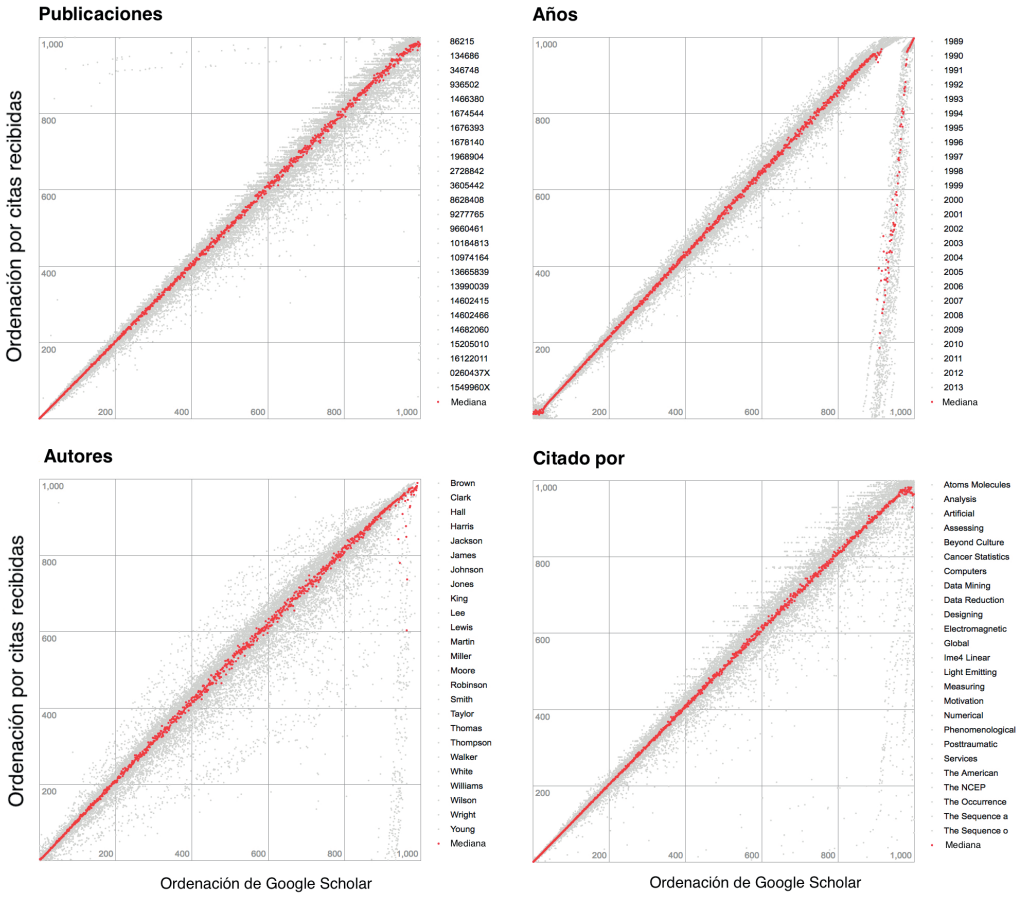


Figure 3. Scatter diagram of the four samples corresponding to the four search types

The software used in the analysis was *R*, version 3.4.0 (*R Development Core Team*, 2017). The confidence intervals were constructed via normal approximation by applying Fisher's transformation using the *R Psych* package (**Revelle**, 2017). Fisher's transformation when applied to Spearman's correlation coefficients is asymptotically normal. Graphs of the confidence intervals were drawn with the *Plotrix* package in *R* (**Lemon**, 2006).

A significance level of alpha = 0.0025 was used to guarantee total confidence of 95% in the interpretation of 25 confidence intervals following Bonferroni's conservative criterion. Setting a low level of significance widens the confidence intervals relative to the intervals obtained with a higher alpha value (e.g., 0.05).

5. Results

Spearman's correlation coefficients between the *Google Scholar* ranking and the citations received rankings were surprisingly high, with values close to or greater than 0.9 (Table 2).

Figures 3 and 4 show that the correlations for the four samples were almost perfect; it was only at position number 900 that we started to find data with no correlation, especially in the case of the search by year. This effect was already detected in previous research (**Martín-Martín** *et al.*, 2014; 2017). The data obtained prove that in these positions, the number of citations received is very small. Therefore, it is highly likely that in these cases, the ranking position is established by other external SEO factors bearing no relation to the citations.

If we had only considered the values up to position 900, the correlation coefficients would have been even greater. Table 3 presents the specific data for each search. The correlation coefficients were all close to 0.9, and they exceeded this value in most cases.

6. Discussion

The correlation coefficients were surprisingly high. In previous applications of reverse engineering to the *Google Search* ranking, coefficients exceeding 0.4 were rarely obtained. Our investigation yielded correlation coefficients greater than 0.9 for all four search types and a global value of 0.9999. Although the four types of searches were completely different from each other (authors, years, publications and "cited by"), they all yielded the same correlation pattern. The correlations by year, however, indicated a greater level of variability than the rest of the searches, while the authors showed greater variability than publications and "cited by" (Figure 4).

Table 3. Spearman's correlation coefficients and confidence limits of the four samples corresponding to the four search types (α = 0.0025)

	Years				Authors				Cited by				Publications			
#	id	rho	L	U	id	rho	L	U	id	rho	L	U	id	rho	L	U
1	1989	0.90	0.88	0.91	Clark	0.97	0.96	0.97	Atoms_molecules	0.99	0.99	0.99	14682060	0.99	0.99	0.99
2	1990	0.87	0.84	0.89	Hall	0.91	0.89	0.93	Analysis	0.99	0.99	0.99	9660461	0.99	0.99	0.99
3	1991	0.87	0.85	0.90	Harris	0.97	0.96	0.97	Artificial	0.98	0.98	0.98	86215	0.99	0.98	0.99
4	1992	0.90	0.88	0.92	Jackson	0.93	0.91	0.94	Assessing	0.99	0.99	0.99	2728842	0.98	0.98	0.99
5	1993	0.89	0.87	0.91	James	0.97	0.96	0.98	Beyond_culture	0.94	0.93	0.95	15205010	0.99	0.99	0.99
6	1994	0.89	0.87	0.91	King	0.96	0.95	0.96	Cancer_statistics	0.99	0.99	0.99	9277765	0.99	0.99	0.99
7	1995	0.88	0.86	0.90	Lee	0.96	0.95	0.97	Computers	0.98	0.97	0.98	936502	0.99	0.99	0.99
8	1996	0.89	0.87	0.91	Lewis	0.97	0.96	0.97	Data_Mining	0.98	0.97	0.98	134686	0.99	0.99	0.99
9	1997	0.89	0.87	0.91	Martin	0.90	0.88	0.92	Data_reduction	0.99	0.99	0.99	1968904	0.99	0.99	0.99
10	1998	0.89	0.87	0.91	Moore	0.96	0.95	0.96	Designing	0.95	0.93	0.95	3605442	0.99	0.99	0.99
11	1999	0.87	0.85	0.89	Robinson	0.95	0.94	0.96	Electromagnetic	0.98	0.98	0.99	10184813	0.99	0.99	0.99
12	2000	0.89	0.87	0.91	Taylor	0.96	0.95	0.96	Global	0.98	0.97	0.98	8628408	0.99	0.99	0.99
13	2001	0.89	0.87	0.91	Thomas	0.90	0.87	0.91	Ime4_Linear	0.95	0.94	0.96	0260437X	0.99	0.99	0.99
14	2002	0.86	0.83	0.88	Thompson	0.94	0.93	0.95	Light_emitting	0.99	0.99	0.99	1674544	0.98	0.98	0.99
15	2003	0.88	0.86	0.90	Walker	0.96	0.95	0.97	Measuring	0.99	0.99	0.99	1549960X	0.99	0.99	0.99
16	2004	0.89	0.86	0.91	White	0.96	0.95	0.97	Motivation	0.90	0.88	0.92	14602466	0.99	0.99	0.99
17	2005	0.92	0.90	0.93	Wright	0.95	0.94	0.96	Numerical	0.97	0.96	0.97	13665839	0.99	0.98	0.99
18	2006	0.83	0.79	0.85	Young	0.94	0.93	0.95	Phenomenological	0.94	0.93	0.95	14602415	0.99	0.99	0.99
19	2007	0.84	0.81	0.86	Wilson	0.97	0.97	0.98	Posttraumatic	0.98	0.97	0.98	16122011	0.99	0.99	0.99
20	2008	0.88	0.86	0.90	Brown	0.97	0.96	0.97	Services	0.94	0.93	0.95	10974164	0.99	0.99	0.99
21	2009	0.90	0.88	0.91	Johnson	0.97	0.96	0.97	The_American	0.99	0.99	0.99	1466380	0.99	0.99	0.99
22	2010	0.88	0.85	0.90	Jones	0.97	0.96	0.97	The_NCEP	0.98	0.98	0.99	1678140	0.99	0.99	0.99
23	2011	0.83	0.80	0.86	Miller	0.97	0.97	0.98	The_occurrence	0.99	0.99	0.99	346748	0.88	0.85	0.90
24	2012	0.89	0.86	0.91	Smith	0.96	0.95	0.97	The_sequence a	0.90	0.88	0.92	1676393	0.99	0.98	0.99
25	2013	0.89	0.86	0.91	Williams	0.96	0.95	0.97	The_sequence o	0.99	0.99	0.99	13990039	0.99	0.99	0.99

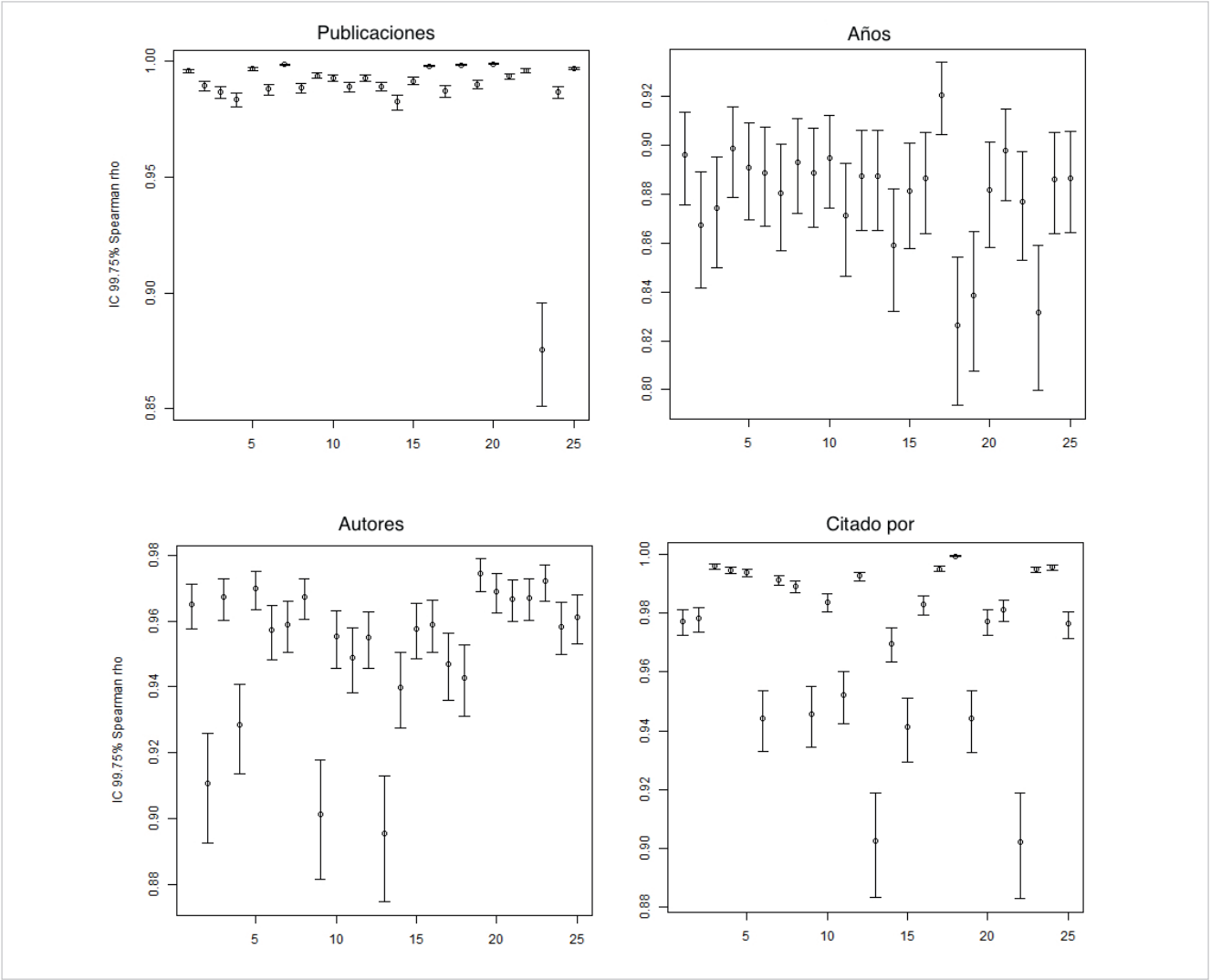


Figure 4. Box plot diagram of the four samples corresponding to the four search types

These results provide conclusive evidence that the number of citations received is of the utmost importance when considering the external SEO features of *Google Scholar*. These conclusions are consistent with those of **Beel** and **Gipp** (2009a) and **Martín-Martín** *et al.* (2014; 2017), who approached the same subject using different methodologies:

- In the first case, we analysed thematic searches with keywords and came to the same conclusion that the citations received may be considered a highly relevant ordering factor. However, the results obtained are not sufficiently clear because they do not analyse the external SEO factors in an isolated fashion.
- In **Martín-Martín** *et al.* (2014; 2017), only searches by year were studied, also yielding correlation coefficients of approximately 0.9.

Our study offers additional evidence because it also analysed those searches by author, publication and "cited by", obtaining the same high correlations in every case. In **Martín-Martín** *et al.* (2017) we can also see how the position in *Google Scholar* is correlated with the number of citations received; therefore, a variable ordinal is being correlated with another more modest quantitative factor. In contrast, our methodological proposal consists of comparing two variable ordinals: the order of *Google Scholar* with the order according to citations (**Beel**; **Gipp** 2009a), based on the average of each sample.

> Academic journals provide instructions to authors about how to improve the *Google Scholar* ranking without giving citations the attention that they merit

Other investigations (**Martín-Martín** *et al.*, 2017; **Moed**; **Bar-Ilan**; **Halevi**, 2016) would appear to point towards the possibility that other external factors, such as the language of the document, the number of versions or the speed of indexing, can influence in the ranking. Nevertheless, with correlations of 0.99 out of 1, we are left with a very slim margin for factors other than citations. The ordering by number of citations is practically the same as the native ordering of *Google Scholar*. The remaining external factors have a merely residual influence, including received links, which are very important in *Google* searches. These factors could play a more important role in documents with few or no citations received, as is the case of items situated from the 900th position onwards.

Guides or instructions to authors from publishers of academic journals generally provide guidelines regarding how to improve the ranking positions of articles in *Google Scholar* (*Elsevier*, 2012; *Wiley*, 2015). These guidelines tend to be contaminated by the ranking of *Google Search*. For instance, it is often stated that the ranking order in *Google Scholar* depends on the publisher, insinuating that some publications have superior positioning to others. The suggestions to authors also affirmed that there are other off-page positioning factors, such as presence in social networks, backlinks or the prestige of an author according to the number of citations received for all of his/her published works (*Google*, 2011).

> "Try to increase the number of citations to improve your paper's ranking on *Google Scholar*"

We found no evidence that any of these affirmations are intervening factors in algorithm ranking. However, they might exert an indirect influence if they enable the published work to be more widely read and thus lead to an increase in the number of citations received.

From this study, we were unable to draw conclusions related to the on-page features regarding how to use keywords in documents to improve positioning. This type of feature has not been studied directly since searches without keywords were used, and internal factors were blocked. Therefore, there is no evidence disputing the recommendations that are typically provided in this context, such as including the most important keywords from the article and their synonyms in the title, subtitles or summary and optimising the number of times that a keyword appears in the article.

7. Conclusions

We developed a new analysis procedure in the context of reverse engineering studies that enabled us to the study off-page SEO features of *Google Scholar* in isolation. Using this new method, we were able to ascertain that citations are the main off-page SEO feature in *Google Scholar*. The statistical results leave no doubt. This new analysis helps to augment the scarce information available about this topic and presents a new method from a different statistical angle that might be of practical use in forthcoming investigations.

By employing reverse engineering, we were able to obtain estimates of the intervening factors in ranking results and their relative importance. Our findings are useful for improving the experimental basis of the ASEO discipline and could provide better recommendations to authors regarding how they might optimise their rankings of published works in *Google Scholar*.

In conclusion, this research, in addition to demonstrating the intrinsic value of ASEO, provided specific recommendations for authors of scientific and academic articles. The primary recommendation developed from this study is to produce good quality articles so that they are widely read and are cited and thus enter the "virtuous circle" in which more citations lead to better rankings, which in turn lead to greater visibility and an increased number of citations (**Martín-Martín** *et al.*, 2016b).

8. Acknowledgment

This work is part of the project *Interactive content and creation in multimedia information communication: Audiences, design, systems and styles*, CSO2012-39518-C04-02, *Spanish Ministry of Economy and Competitiveness* (*Mineco/Feder*).

9. References

Aguillo, Isidro F. (2012). "Is Google Scholar useful for bibliometrics? A webometric analysis". *Scientometrics*, v. 91, n. 2, pp. 343-351.
https://goo.gl/nYBmZb

Beel, Joeran; **Gipp, Bela** (2009a). "Google Scholar's ranking algorithm: An introductory overview". In: *Procs of the 12th intl conf on scientometrics and informetrics*, *ISSI'09*, pp. 230-241.
https://goo.gl/c8a6YU

Beel, Joeran; **Gipp, Bela** (2009b). "Google Scholar's ranking algorithm: the impact of articles' age (an empirical study)". In: *6th intl conf on information technology: New generations*, *ITNG'09*, pp. 160-164.
https://goo.gl/cfV2my
https://doi.org/10.1109/ITNG.2009.317

Beel, Joeran; **Gipp, Bela** (2009c). "Google Scholar's ranking algorithm: the impact of citation counts (an empirical study)". In: *3rd intl conf on research challenges in information science*, *RCIS 2009*, pp. 439-446.
https://www.gipp.com/wp-content/papercite-data/pdf/beel09a.pdf
https://doi.org/10.1109/RCIS.2009.5089308

Beel, Joeran; **Gipp, Bela** (2010). "Academic search engine spam and Google Scholar's resilience against it". *The journal of electronic publishing*, v. 13, n. 3, pp. 1-28.
https://doi.org/10.3998/3336451.0013.305

Beel, Joeran; **Gipp, Bela**; **Wilde, Erik** (2010). "Academic search engine optimization (ASEO). Optimizing scholarly literature for Google Scholar & co". *Journal of scholarly publishing*, v. 41, n. 2, pp. 176-190.
https://docear.org/papers/Academic%20Search%20Engine%20Optimization%20(ASEO)%20--%20preprint.pdf
https://doi.org/10.3138/jsp.41.2.176

Codina, Lluís (2017). "SEO académico: definición, componentes y guía de herramientas". *Lluís Codina*, 17 noviembre.
https://www.lluiscodina.com/seo-academico-guia

De-Groote, Sandra L.; **Raszewski, Rebecca** (2012). "Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing". *Nursing outlook*, v. 60, n. 6, pp. 391-400.
https://doi.org/10.1016/j.outlook.2012.04.007

De-Winter, Joost C. F.; **Zadpoor, Amir A.**; **Dodou, Dimitra** (2014). "The expansion of Google Scholar versus Web of Science: A longitudinal study". *Scientometrics,* v. 98, n. 2, pp. 1547-1565.
https://doi.org/10.1007/s11192-013-1089-2

Delgado-López-Cózar, Emilio; **Robinson-García, Nicolás**; **Torres-Salinas, Daniel** (2012). *Manipular Google Scholar Citations y Google Scholar Metrics: simple, sencillo y tentador*. *EC3 working papers.* Granada: Universidad De Granada.
http://hdl.handle.net/10481/20469

Delgado-López-Cózar, Emilio; **Robinson-García, Nicolás**; **Torres-Salinas, Daniel** (2014). "The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators". *Journal of the Association for Information Science and Technology*, v. 65, n. 3, pp. 446-454.
https://arxiv.org/abs/1309.2413
https://doi.org/10.1002/asi.23056

Elsevier (2012). "Get found: optimize your research articles for search engines". *Elsevier*, 6 Nov.
https://www.elsevier.com/connect/get-found-optimize-your-research-articles-for-search-engines

Emerald Publishing Limited (2017). "How to... disseminate your work". *Emerald Publishing*.
http://www.emeraldgrouppublishing.com/authors/guides/promote/disseminate.htm

Enge, Eric; **Spencer, Stephan**; **Stricchiola, Jessie** (2015). *The art of SEO: mastering search engine optimization.* Sebastopol CA: O'Reilly Media. ISBN: 978 1 491903643

Farhadi, Hadi; **Salehi, Hadi**; **Yunus, Melor**; **Aghaei-Chadegani, Arezoo**; **Farhadi, Maryam**; **Fooladi, Masood**; **Ale-Ebrahim, Nader** (2013). "Does it matter which citation tool is used to compare the h-index of a group of highly cited researchers?". *Australian journal of basic and applied sciences*, v. 7, n. 4, pp. 198-202.
https://ssrn.com/abstract=2259614

Gielen, Matt; **Rosen, Jeremy** (2016). "Reverse engineering the YouTube algorithm: Part I". *Tubefilter.com*, 23 June.
http://www.tubefilter.com/2016/06/23/reverse-engineering-youtube-algorithm

Giustini, Dean; **Boulos, Maged N. K.** (2013). "Google Scholar is not enough to be used alone for systematic reviews". *Online journal of public health informatics*, v. 5, n. 2, pp. 1-9.
https://doi.org/10.5210/ojphi.v5i2.4623

Google (2011). "About Google Scholar". *Google Scholar*.
http://scholar.google.com/intl/en/scholar/about.html

Google (2017). "How Google search works. Learn how Google discovers, crawls, and serves web pages". *Google. Search console help*.
https://support.google.com/webmasters/answer/70897?hl=en

Harzing, Anne-Wil (2011). *The publish or perish book: your guide to effective and responsible citation analysis*. Melbourne, Australia: Tarma Software Research Pty Ltd. ISBN: 978 1 60752 120 4
https://harzing.com/publications/publish-or-perish-book/pdf

Harzing, Anne-Wil (2013). "A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel prize winners". *Scientometrics*, v. 94, n. 3, pp. 1057-1075.
https://doi.org/10.1007/s11192-012-0777-7

Harzing, Anne-Wil (2014). "A longitudinal study of Google Scholar coverage between 2012 and 2013". *Scientometrics*, v. 98, n. 1, pp. 565-575.
https://harzing.com/download/gs_coverage.pdf
https://doi.org/10.1007/s11192-013-0975-y

Jacsó, Péter (2008a). "Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for FW Lancaster". *Library trends*, v. 56, n. 4, pp. 784-815.
https://goo.gl/Sj5QBr
https://doi.org/10.1353/lib.0.0011

Jacsó, Péter (2008b). "The pros and cons of computing the h-index using Google Scholar". *Online information review*, v. 32, n. 3, pp. 437-452.
https://goo.gl/Hz8ABM
https://doi.org/10.1108/14684520810889718

Jacsó, Péter (2009). "Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software". *Online information review*, v. 33, n. 6, pp. 1189-1200.
https://doi.org/10.1108/14684520911011070

Jacsó, Péter (2012). "Using Google Scholar for journal impact factors and the h-index in nationwide publishing assessments in academia –siren songs and air-raid sirens". *Online information review*, v. 36, n. 3, pp. 462-478.
https://goo.gl/FfrDgt
https://doi.org/10.1108/14684521211241503

Jamali, Hamid R.; **Asadi, Saeid** (2010). "Google and the scholar: the role of Google in scientists' information-seeking behaviour". *Online information review*, v. 34, n. 2, pp. 282-294.
https://goo.gl/cZrMwi
https://doi.org/10.1108/14684521011036990

Jamali, Hamid R.; **Nabavi, Majid** (2015). "Open access and sources of full-text articles in Google Scholar in different subject fields". *Scientometrics*, v. 105, n. 3, pp. 1635-1651.
https://doi.org/10.1007/s11192-015-1642-2

Lemon, Jim (2006). "Plotrix: a package in the red light district of R". *R-News*, v. 6, n. 4, pp. 8-12.
https://www.researchgate.net/publication/260171541_Plotrix_A_package_in_the_red_light_district_of_R

Localseoguide (2016). "Local SEO ranking factors study 2016". *Localseoguide*.
http://www.localseoguide.com/guides/2016-local-seo-ranking-factors

Maciá-Domene, Fernando (2015). *SEO: técnicas avanzadas.* Barcelona: Anaya. ISBN: 978 84 41537309

Marcos, Mari-Carmen; **González-Caro, Cristina** (2010). "Comportamiento de los usuarios en la página de resultados de los buscadores. Un estudio basado en *eye tracking*". *El profesional de la información*, v. 19, n. 4, pp. 348-358.
https://doi.org/10.3145/epi.2010.jul.03

Martín-Martín, Alberto; **Ayllón, Juan-Manuel**; **Orduña-Malea, Enrique**; **Delgado-López-Cózar, Emilio** (2016a). *Google Scholar Metrics released: a matter of languages… and something else.* Granada: Universidad de Granada.
https://arxiv.org/abs/1607.06260v1

Martín-Martín, Alberto; **Orduña-Malea, Enrique**; **Ayllón, Juan-Manuel**; **Delgado-López-Cózar, Emilio** (2016b). "Back

to the past: On the shoulders of an academic search engine giant". *Scientometrics*, v. 107, n. 3, pp. 1477-1487.
https://arxiv.org/abs/1603.09111
https://doi.org/10.1007/s11192-016-1917-2

Martín-Martín, Alberto; **Orduña-Malea, Enrique**; **Ayllón, Juan-Manuel**; **Delgado-López-Cózar, Emilio** (2014). *Does Google Scholar contain all highly-cited documents (1950-2013)? EC3 working papers.* Granada: Universidad de Granada.
https://arxiv.org/abs/1410.8464

Martín-Martín, Alberto; **Orduña-Malea, Enrique**; **Harzing, Anne-Wil**; **Delgado-López-Cózar, Emilio** (2017). "Can we use Google Scholar to identify highly-cited documents?". *Journal of informetrics*, v. 11, n. 1, pp. 152-163.
https://arxiv.org/abs/1804.10439
https://doi.org/10.1016/j.joi.2016.11.008

Mayr, Philipp; **Walter, Anne-Kathrin** (2007). "An exploratory study of Google Scholar". *Online information review*, v. 31, n. 6, pp. 814-830.
https://arxiv.org/pdf/0707.3575.pdf
https://doi.org/10.1108/14684520710841784

Merton, Robert K. (1968). "The Matthew effect in science". *Science*, v. 159, n. 3810, pp. 56-63.
https://goo.gl/Zcpqcc
https://doi.org/10.1126/science.159.3810.56

Moed, Henk F.; **Bar-Ilan, Judit**; **Halevi, Gali** (2016). "A new methodology for comparing Google Scholar and Scopus". *Journal of informetrics*, v. 10, n. 2, pp. 533-551.
https://arxiv.org/abs/1512.05741
https://doi.org/10.1016/j.joi.2016.04.017

MOZ (2015). *Search engine ranking factors 2015*.
https://moz.com/search-ranking-factors/correlations

MOZ (2017). *Google algorithm change history*.
https://moz.com/google-algorithm-change

Muñoz-Martín, Beatriz (2015). "Incrementa el impacto de tus artículos y blogs: de la invisibilidad a la visibilidad". *Revista de la Sociedad Otorrinolaringológica de Castilla y León, Cantabria y La Rioja*, v. 6, n. Suppl. 4, pp. 6-32.
http://hdl.handle.net/10366/126907

Orduña-Malea, Enrique; **Ayllón, Juan-Manuel**; **Martín-Martín, Alberto**; **Delgado-López-Cózar, Emilio** (2014). *About the size of Google Scholar: playing the numbers. EC3 working papers.* Granada: Universidad de Granada.
https://arxiv.org/abs/1407.6239

Orduña-Malea, Enrique; **Ayllón, Juan-Manuel**; **Martín-Martín, Alberto**; **Delgado-López-Cózar, Emilio** (2015). "Methods for estimating the size of Google Scholar". *Scientometrics*, v. 104, n. 3, pp. 931-949.
https://doi.org/10.1007/s11192-015-1614-6

Orduña-Malea, Enrique; **Martín-Martín, Alberto**; **Ayllón Juan-Manuel**; **Delgado-López-Cózar, Emilio** (2016). *La revolución Google Scholar: destapando la caja de Pandora académica.* Granada: Editorial Universidad de Granada. ISBN: 978 84 33859419
https://goo.gl/3oUGKQ

Pedersen, Lee A.; **Arendt, Julie** (2014). "Decrease in free computer science papers found through Google Scholar". *Online information review*, v. 38, n. 3, pp. 348-361.
https://goo.gl/ngmZ1Q
https://doi.org/10.1108/OIR-07-2013-0159

R Development Core Team (2008). *R: a language and environment for statistical computing*.
http://softlibre.unizar.es/manuales/aplicaciones/r/fullrefman.pdf
https://www.R-project.org

Ratcliff, Christopher (2016). "WebPromo's Q&A with Google's Andrey Lipattsev, search engine watch". *Search*
https://searchenginewatch.com/2016/04/06/webpromos-qa-with-googles-andrey-lipattsev-transcript

Revelle, William (2017). *Psych: procedures for personality and psychological research*. Northwestern University.
https://CRAN.R-project.org/package=psych

Schwartz, Barry (2016). "Now we know: Here are Google's top 3 search ranking factors". *Search engine land*, 24 March.
http://searchengineland.com/now-know-googles-top-three-search-ranking-factors-245882

Searchmetrics (2016). "Rebooting ranking factors". *Searchmetrics.*
http://www.searchmetrics.com/knowledge-base/ranking-factors

Torres-Salinas, Daniel; **Ruiz-Pérez, Rafael**; **Delgado-López-Cózar, Emilio** (2009). "Google scholar como herramienta para la evaluación científica". *El profesional de la información*, v. 18, n. 5, pp. 501-510.
https://doi.org/10.3145/epi.2009.sep.03

Van-Aalst, Jan (2010). "Using Google Scholar to estimate the impact of journal articles in education". *Educational researcher*, v. 39, n. 5, pp. 387-400.
https://goo.gl/p1mDBi

Van-der-Graaf, Peter (2012). "Reverse engineering search engine algorithms is getting harder". *Search engine watch*, 7 June.
https://searchenginewatch.com/sew/how-to/2182553/reverse-engineering-search-engine-algorithms-getting-harder

Walters, William H. (2008). "Google Scholar search performance: Comparative recall and precision". *Portal: Libraries and the academy*, v. 9, n. 1, pp. 5-24.
http://cdm15970.contentdm.oclc.org/utils/getfile/collection/p15970coll1/id/83/filename/84.pdf
https://doi.org/10.1353/pla.0.0034

Wiley (2015). *Writing for SEO*.
https://authorservices.wiley.com/author-resources/Journal-Authors/Prepare/writing-for-seo.html