A Collection of Samples for Research in Google: Design and Application of a Sample Selection Method: Results and Problems of Research

Joan Pedro

Abstract

This article examines the use and validity of Google's search engine for the collection of corpuses of materials for research. To this end, it develops two interrelated themes. In the first section, a methodology is developed which is designed to identify universes in Google that meet the criteria and parameters required by an academic study. This methodology makes use of the search engine's own logic and is applicable to most on-line document searches. The second section discusses the limitations and skewing of results arising from Google's mode of operation which have an impact on the scientific validity of the universes it generates. This part focuses on the completeness and representativeness of the Google universes with regards to the full range of contents available on the Internet.

Keywords: Internet, Google, on-line research, methodology, skewing

Introduction: The use of search engines for the collection of research materials

Search engines have opened up a wide range of possibilities that are worthy of consideration given their impact on expanding and improving research methods. At the same time, the use of search engines has produced specific problems related to the validity of the corpus of materials they generate.

Google and other search engines have made the biggest contribution to studies of sources in the history of research. They offer the most flexible method in existence for the selection of information related to any topic of study, making materials available for analysis with an immediacy, abundance and variety that once would have been unthinkable.

Google is, in itself, a generator of researchable universes, as it uses a complex algorithm to select the materials available according to the criteria established by the researcher. It is worth stressing from the outset that the results are determined by the query indicated, but in accordance with the internal criteria applied by each search engine.

As explained later, no program covers all of the content existing on-line. This means that, in general, Google and other search engines will not bring up all of the available web pages related to a given query; it will produce a sample. To put it simply, when using Google, the universe being explored is the Google universe, not the Internet universe.

Objectives

This article explores two closely interrelated issues. Firstly, it proposes a methodology designed and tested to identify research universes and optimise Google searches. Google is the search engine that works best for the compilation of universes, for example, by establishing parameters that would be necessary in many content studies of on-line materials, as in the case of the research from which the models and results presented here are taken.

Secondly, the way in which Google operates needs to be understood and taken into account in studies with scientific methodologies, because they affect the representativeness and relevance of the sample. An analysis of this question is developed in the second part of the article.

Methodological Considerations

Searching for universes in Google is one of the tasks performed by the Group 'Identidades Sociales y Comunicación' in the context of a research project financed by the Universidad Complutense de Madrid and Banco de Santander, under the title 'Regulated and Non-regulated Communications Studies offered on the Internet: Repercussions on the Transformation of University Studies'. The Group and this study are directed by Olivia Velarde of the Universidad Complutense de Madrid. It is a study of the orientations being given to communications studies, and forms part of a series of R&D studies being conducted by the Group on 'Social Communication in the Age of Globalisation'. The models for working with Internet search engines in order to establish research universes and select samples are an application of the logical methodologies designed by Manuel Martín Serrano (1974, 1977, 1986, 2008) for content analysis.

From the range of training programs appearing on the Internet, this study has as a corpus those programs that provide training related to communications (media, journalism, advertising, and other programs that are identified as communications studies). In addition, it takes into account the type of program offered (courses, bachelor's degrees, post-graduate studies, masters, workshops, etc., public and private, regulated and non-regulated, on-line and/or in-class).

The present article explains how it was delimited and found the universe of all the training programs offered in Spain providing training related to communications that appear in Google search results.

On the other hand, part two of the article draws on the research findings, as well as on research on the structure of the Internet and on the functioning of Google itself, especially, but not limited to the fields of network topology and search engine optimisation.

Criteria for delimiting the universe of the search

The following parameters were established to specify the characteristics of the universe of training programs related to communications on the Internet:

- Spatio-temporal parameters
- Key word parameters:
    - Parameters of the universe of training programs
    - Parameters of samples according to use of the Internet
    - Filtering of the samples
- Limitation of page titles only ('allintitle')

Spatio-temporal parameters

Spatial parameters: *Pages from Spain*. The button to select this option is located in the left margin of the results page.

Temporal parameters: *Past year.* This option is found in the 'More search tools' function located in the left margin of the results page. This function also offers the possibility of searches for the past hour, past 24 hours, past month or within a specified period (in years).

Google also offers the possibility of doing a search starting on the month, week or day indicated by the user up to present time; after completing a key word search, the researcher must add to the end of the Google search address (URL) the characters '&as_qdr=', followed by the letter 'd' to specify the days, 'w' to specify the weeks, 'm' in the case of months or 'y' for years. After the letter, the researcher must add the number to which the letter refers. For example:

- &as_qdr=d3 will offer results for the past three days.
- &as_qdr=w5 will offer results for the past five weeks.

Key word parameters

Google accepts up to 32 key words in a search and will ignore any additional words. In this model, four different types of key terms have been established to be able to find the corpus of relevant programs in Google. This design makes use of the search engine's own logic and is a method applicable to most on-line document searches. The parameters and the terms used in this research are indicated below. The words are shown in the language in which the search was conducted (Spanish). The English translation of these words is as follows: Type A: course, masters, undergraduate, postgraduate, doctorate, bachelor, diploma, seminar, "professional training", examinations, workshop; Type B: communication, information, journalism, advertising, "public relations", marketing; Type C: online, on-line, virtual, digital, Internet. Type D: chat, forum, blog, news.

It is important to note that the order in which the key words are placed affects the number, content and order of the search results. This aspect is discussed further in the article.

Finally, even if the instructions for filtering the selection of materials have been followed correctly, a comparison of the results is necessary to eliminate results that are not relevant or that are repeated.

Parameters of the universe of training programs:

Type A words (how it is classified): curso, máster, master, grado, postgrado, posgrado, doctorado, licenciatura, diplomatura, seminario, "formación profesional", oposiciones, taller.

Type B words (how it is identified): comunicación, información, periodismo, publicidad, "relaciones públicas", marketing.

Parameters of samples according to use of the Internet:

Type C words: Two exclusive samples will be obtained in this universe according to whether or not the programs have an on-line dimension. To do this, two queries are entered. In one, the search is delimited in order to display only those programs that include the words online, on-line, virtual, digital, or internet. In this way, programs delivered on-line are found, as well as programs whose content makes reference to the on-line world, even if they are delivered in face-to-face classes. In the second search, the same terms are added, but each one is preceded by the symbol [-], so that Google will exclude pages containing those terms. In this way, on-site programs that make no reference to the Internet are obtained.

Filtering of the samples:

Type D words: Google is instructed to omit results that contain the words chat, foro, blog and noticias. To do this, the symbol [-] is entered with no space before each term to be excluded. In this way, materials originating from sources that are not relevant to this study are avoided.

Limitation of page titles only ('allintitle')

In many content studies it is useful to work with titles or labels for the selection of corpus materials. In the study referred to here, the search for materials for content analysis using key words is limited to the page title only. This selection is obtained by using the Google 'allintitle' function. This function must be entered at the beginning of the keyword formula followed by a colon and then, without leaving a space, by the rest of the terms (e.g. allintitle:master comunicación).

Formulae for searches with Google's Boolean operators and results

To perform a search, key terms are interrelated using the Boolean operators AND, OR and NOT offered by Google. AND indicates a term that has to appear *together with* another term. It is not necessary to actually enter the word, as leaving a space between terms has the same effect. The command OR (which may be substituted using the symbol [|]) is not used by Google as a criterion of exclusion. For example, it does not search for X or Y, but searches for *either* X or Y, or *both* together; in other words, it searches for pages that contain one and both of the two key words. NOT serves to exclude pages that contain the terms that a user seeks to avoid. It can be substituted by the symbol [-].

In the study discussed here, the aim was to develop a formula that would:

1. allow the location of websites referring to the online world whose titles include a reference to any type of program (masters, seminar, bachelor's, etc.) or *the combination* of two or more Type A terms, *together with* any type of communications field (journalism, communication, marketing, etc.) or *the combination* of two or more Type B terms, *together with* any term or combination of terms referring to the on-line world (online, virtual, on-line, etc.) or Type C terms, *except for* websites with undesired terms in their titles (chat, forum, blog) or Type D terms. Therefore, the entries sought should contain one or more Type A words plus one or more Type B words plus one or more Type C words in the title, provided that the title does not contain any [-] Type D words. The formula for this series of parameters is as follows:

    allintitle:(curso OR máster OR master OR grado OR postgrado OR posgrado OR doctorado OR licenciatura OR diplomatura OR seminario OR "formación profesional" OR oposiciones OR taller) (comunicación OR información OR periodismo OR publicidad OR "relaciones públicas" OR marketing) (online OR on-line OR virtual OR digital OR internet) -chat -foro -blog -noticias

As can be seen, the first set of parentheses refers to Type A words, the second to Type B and the third to Type C, while the words appearing after the third set of parentheses preceded by the symbol [-] refer to the Type D filters. With this formula, a total of 323 training programs are obtained.

2. allow the location of the websites of training programs in communications that do not make any reference to the on-line world. With a simple modification, a sample of websites can be obtained whose titles make no reference to the on-line world. To do this, the same general formula is used, but the Type C terms (online, on-line, virtual, digital, internet) are located as filters [-] at the end of the formula:

> allintitle:(curso OR máster OR master OR grado OR postgrado OR posgrado OR doctorado OR licenciatura OR diplomatura OR seminario OR "formación profesional" OR oposiciones OR taller) (comunicación OR información OR periodismo OR publicidad OR "relaciones públicas" OR marketing) -chat -foro -blog -noticias -online -on-line -virtual -digital -internet

With this formula, a total of 350 results are obtained.

In short, when searches are conducted applying the formula and criteria indicated, a universe of 673 training programs is obtained, made up of 323 referring to the on-line world and 350 with no such reference. In this way, exclusive samples can be obtained from the same universe, with the possibility of comparing differences and similarities.

The different types of result totals offered by Google

For a single search, Google offers three different figures referring to the total number of results found. The meaning of each figure and the criteria for selecting one of the options are explained below.

First figure: Total hits

This number is visible below the search engine interface. It refers to the *total number of hits*, but is generally irrelevant to the research for two reasons. First of all, it is an approximate figure (Ayuda para Webmasters de Google', n.d.). Secondly, this figure cannot be used to estimate the size of the universe that may be accessed through the search engine because Google only displays a maximum of 1000 results, even if the total number of hits indicated is greater.

Second figure: Unique results

This figure is the actual number of pages that Google can display. To see this figure, the user must go to the last page of results, at the bottom of which there will appear some text in italics. In order to make this step quicker it is advisable to configure the search to show 100 results per page. This figure counts only those results that Google considers *unique*. In this search, Google's operations are geared towards omitting repetitions, spam and other entries that it determines are *very similar* to those displayed (although it is important to remember that the search engine has most probably found a certain number of repetitions and irrelevant entries). Moreover, in this selection the number of pages returned from a single domain is also limited.

This was the option used in this study because the results constitute the universe that best meets the criteria of relevance in most studies, and that poses the fewest problems. It is to be expected, however, that this will result in the loss of entries which Google has omitted for being deemed similar to others that it has selected and which might nevertheless be relevant.

Third figure: Very similar results that have been omitted

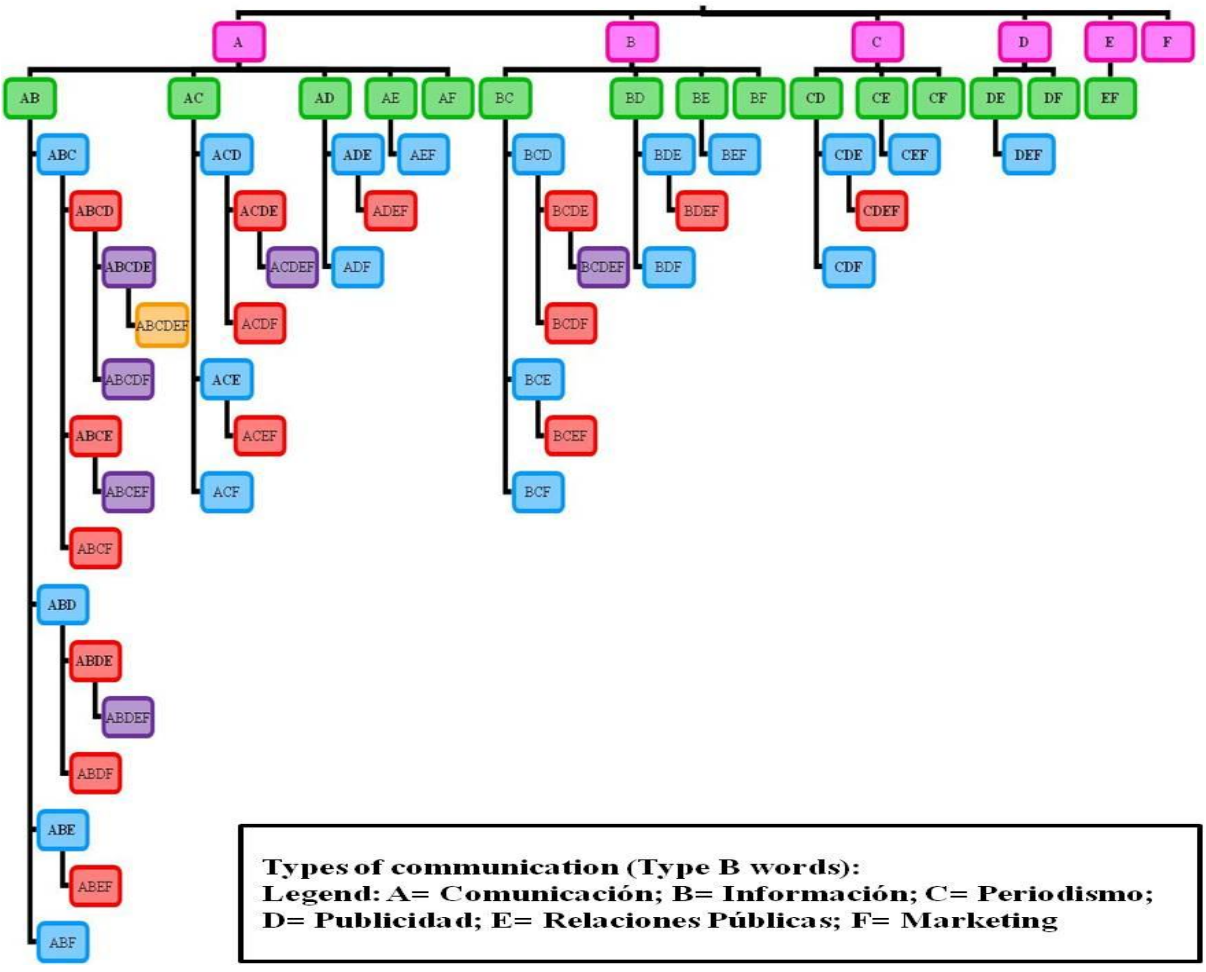
A third figure can be obtained using the function *repeat the search with the omitted results included*, which appears at the bottom of the last page of results in Google. As mentioned above, these omitted results consist of pages that Google deems the same or *very similar* to pages it has already provided. The researcher should consider in each case whether too many repetitions, irrelevant entries or other types of noise have been produced. In this study, it was established that the level of noise was too high.

A combined model to find all possible sub-universes based on disaggregated searches

There are several methods for identifying the sub-universes arising from all the possible combinations of the search terms. The simplest and most efficient method consists in programming a model of search structures that will work with all the possible combinations of key words without producing repetitions or omissions. This model is one of the variants of structure analysis based on the formal logic proposed by Martín Serrano (1974, 1978a, 2008), which have subsequently been used in a wide range of studies by their author and by many other researchers.

Information on the scientific significance and dissemination of these methodologies can be found in the special issue of the journal *Mediaciones sociales* (2007) dedicated to the topic. An outline of the logical bases of the model being applied here and its implementation are beyond the scope of this article, but details can be found in various publications, among which the most relevant to this study is the article "Un método lógico para analizar los significados" (Martín Serrano, 1978b).The following figure represents the range of combinations and the way in which the combinations are derived from others. Each letter corresponds to one of the types of communication (Type B words). Each box refers to a combination of letters included in the search with a specific formula, using the key terms that correspond to each letter.

Figure 1: LOGICAL MODEL TO OBTAIN THE FULL RANGE OF COMBINATIONS OF ELEMENTS WITHOUT REPETITION



This hierarchical diagram covers the whole range of different search structures that can be differentiated using each key word and the complete range of its combinations with the other words. The researcher first performs a search for materials in which a single element appears (first for A, then for B, then for C, D, E and F). This would be followed by pair searches (A with B, A with C, and so on successively through all the green boxes); searches in threes (A with B and C; A with C and D, and so on successively through all the blue boxes). The same procedure applies for combinations of four, five and six elements. Each of the elements that DO NOT APPEAR in a box must be included as a filter **[-]** in the search formula. By way of example, the following is the formula required for the ADF box, which provides a sample that works simultaneously with 'comunicación', 'publicidad' and 'marketing':

allintitle:(curso OR máster OR master OR grado OR postgrado OR posgrado OR doctorado OR licenciatura OR diplomatura OR seminario OR "formación profesional" OR oposiciones OR taller) (comunicación publicidad marketing) (online OR "on-line" OR virtual OR digital OR Internet) -chat -foro -blog -noticias -"relaciones públicas" -periodismo -información

It should be noted that the total universe obtained through this method of disaggregated searches does not exactly match the one obtained with a single combined search. In the study on which this article is based, the disaggregated search produced 100 more results than the aggregated search; it is therefore more complete.

Problems and Skewed Samples in Google Searches

Google's search engine is a tool with great advantages when conducting searches aimed at the identification of universes. However, it is important to consider certain structural limitations and operating procedures of the search engine that hinder the collection of the desired information and affect the representativeness of research corpuses.

The selection and ordering of the results responds to an indexing process and an algorithm that the company has developed. Although the algorithm is kept secret and is prone to alteration, some variables have been made public which may help to understand some aspects of how Google produces its universes (Cf. Aubuchon, 2009; Fleischner, 2009; Google, 2010; Google Centro para Webmasters, n.d.; Google Corporate Information, n.d.; Schwartz, 2008; Smarty, 2009). Moreover, academic research, which will be cited, on the topology of the Internet and the functioning and results offered by Google itself provide key findings.

A necessary criterion to consider in the generation of results is the key words used in a search. Google focuses on what it calls Density. This means that Google assigns a percentage value to each page depending on the number of times that the key words appear in different parts of the page. The information available indicates that this percentage is derived from a weighting of the number of times that the terms appear in locations such as the URL, page title, description, headings, names of the links, words in bold, and alternative texts such as images, and their proximity (Aubuchon, 2009; Google, 2010).

This criterion of Density clearly meets certain values of relevance and representativeness. However, among the criteria used by Google, there are some important factors that skew the formation and ordering of universes, which need to be examined. Indeed, in our search, Google offered some results which were repeated, some not pertinent (e.g. specialised search engines), and, more importantly, some notable exclusions of relevant communications programs. A review of specialised literature suggests that one of the most relevant aspects is the fact that the selection and ordering of the results respond to hierarchical criteria which tend to favour sites belonging to established, dominant institutions, at the expense of new and less well-established sites, and thus for innovation and diversity. The researcher does not have access to some sources that are necessary for the search to be complete, and even indispensable in certain research projects.

Based on the research findings and on specialised literature, the following sections examine several types of limitations and problems posed by the search engine.

Limitation of the size of searches to 1000 results

Google currently returns a maximum of 1000 results, which means that the universe of a study can never be greater than this figure if obtained using a single search. This means that when the total number of relevant materials available on-line is greater, this universe will be incomplete.

The order of the key words affects the search results

This study has identified differences in results by searching with the same terms of the general formula for the on-line sample, but randomly modifying the order in which the words are positioned. This produced differences to the order of 10 results over the 323 original results. Google responded to a query on this issue, confirming that, indeed, the order of key words influences both the number and the content of results. To improve the representativeness of the corpus of data, it may be necessary to work with a 'sample of samples' obtained by varying the order of the search terms.

It is not possible to compile a universe of topics that require the use of more than 32 terms

As mentioned above, the search formula cannot include more than 32 terms. This is a considerable number for differentiating universes; nevertheless, when it becomes necessary to divide the sub-samples into variables and categories this number of differentiations is no longer sufficient. This limitation hinders the use of structural and discriminative methodologies for content analysis. As a result, Google is an operator that could rarely substitute the use of traditional calculation models and programs for processing results.

Capacity of the search engine to alter the ranking algorithm

On a more general level, an important factor to note is that Google can modify the search results by altering its algorithm. It has previously been possible to confirm that Google has modified its algorithm to favour or hinder certain websites. It is well-known that the company has been subjected to censorship in China so that its search engine would only display websites permitted by the Chinese government. Moreover, according to a study by Edelman & Zittrain (2002), Google has excluded 113 websites (totally or partially) that appear in *Google.com* from the French and German versions of Google. Companies such as Foundem, Ejustice and Ciao! (owned by Microsoft) have filed the most recent charges of manipulation against Google. From the perspective of research, the representativeness and reliability of the samples obtained from this source cannot be assured as long as their decisions continue to be discretionary and are not made public. In addition to algorithm alterations, it is also worth noting certain standard procedures and criteria used by Google for website selection and ordering. As the information is indexed and distributed/organised hierarchically, 'equality of opportunity' does not exist for all materials. The exhaustiveness and representativeness of the corpus will be compromised by the operating codes of the search engine.

In Google, as in all other generalist search engines, there is a skewing which undermines the equiprobability of results. This skewing has an impact on:

- The sources that users can access via the searches.
- The sources that users will not be able to access because they do not appear.
- The sources that they are more likely to access because they appear in the first positions.
- The sources that are more difficult to access because they appear in less visible positions.

This is explored in the following sections.

Limitations of indexation: Lack of exhaustiveness and representativeness

As indicated earlier, the results provided by Google on a particular topic is only one part of the total universe available on the Internet on that topic. One reason for it is that Google does not index (include in its databases) all web materials and, therefore, not all possible records can be included in its results. This means that the indexes or databases do not meet the criterion of completeness. In addition, they also fail to meet the criteria of representativeness, as the materials are not selected for the databases randomly, but based on their links. The chance that a web page has of being included in Google's indexes depends on how well connected (linked) it is. This procedure eliminates pages that do not receive links and hinders the indexing of sites that receive only a few. However, connection indexes are not always indicators of the relevance of the information. In fact, sites that have not been linked could be an important part of research corpuses.

It is practically impossible to know exactly how many pages exist on the web and how many are covered by search engines. The studies conducted differ substantially in the figures given (and the methodologies used), but the conclusion that can be drawn from these studies is that search engines do not index much of the web content

available on the Internet. This is so, firstly, because search engines do not index all the materials in what specialists call the 'visible web', i.e. materials that are indexable by search engines (Barfourosh et al, 2002; Gulli & Signorini, 2005; Lawrence & Giles, 1999), and secondly, because there is another part of the web, called the 'deep web', 'invisible web' or 'dark matter', which contains material that search engines do not access, or have only limited access to, for various technical and human reasons, and which, according to some estimates, is considerably larger than the 'visible web' (Bergman, 2001; He et al, 2007; Sherman & Price, 2001; cf. Sweeny et al, 2010; cf. Wouters et al, 2006). Google has progressively implemented procedures in its system that enable the indexing of deep web content as well (cf. Madhavan et al, 2008), but there is still a lot of content that remains invisible. While it must be stressed that much more work is needed to reach a clear understanding about the deep web and search engine indexing, He et al (2005) found that Google indexed 32% of material from a sample of the deep web.

In terms of representativeness, the Google indexing process proves problematic due to the structure of Internet links themselves. Academic debates on networks with a complex topology like the Internet are still taking place, but there is evidence that rather than constituting a random network of nodules, the Internet is a scale-free network with a power-law distribution. This was first found by Barabási & Albert (1999) who discovered that a few nodes, identified as 'hubs', receive most of the links, while the majority of sites receive very few links, and that new nodes incorporated into the network are linked preferentially to the most connected nodes due to a process of 'preferential attachment'. In this way, nodes with a large quantity of connections tend to accumulate more links quickly, while those that possess few links are rarely the source of new ones.

Since the publication of their article, a considerable number of studies have aimed at refining and improving the understanding of the structure of the Internet. The authors themselves have modified their model based on the observations of Huberman & Adamic (1999) to demonstrate that the *average* number of connections of the oldest nodes in relation to the *average* number of links of new nodes follows a power-law, but that there is also *an intrinsic growth factor for each node* (cf. Benkler, 2006). As Benkler (2006, p. 247) suggests, this modification is important because it specifies that not all new sites are condemned to a marginal position in terms of links, but that the chance of their becoming linked is much lower *on average* than it is for the dominant sites. In other words, there is also room for rapid growth of links to new nodes. In this regard, the NEC model (Pennock et. al, 2002) is of particular interest. The authors of this model sustain that the distribution of connectivity on the Net is close to a pure power-law, but that the distribution of links varies according to the fields or categories to which the sites belong. Their model combines 'preferential attachment' with 'uniform attachment' (unimodal distribution on the logarithmic scale), which has enabled the authors to quantify the degree to which the 'rich' (heavily linked) sites become richer, and the chances of new sites being able to compete according to the field in which they operate.

The debate and the most developed and complex conceptualisations on the structure of on-line links (cf. Benkler, 2006) are beyond the scope of this article, so only a summary is offered here of the findings that are the most relevant to this study. According to the specialist Benkler (2006),

> [theoretical and empirical literature] has consistently shown that the number of links into and out of Web sites follows power laws and that the exponent (the exponential factor that determines that the drop-off between the most linked-to site and the second most linked-to site, and the third, and so on, will be so dramatically rapid, and how rapid it is) for inlinks is roughly 2.1 and for outlinks 2.7. (pp. 244 – 245)

All of this shows that there is an imbalance in power relations, whereby a minority of websites have a much greater influence than the majority, but that there are variable and limited possibilities of being able to compete with the dominant sites. This inequality intrinsic to the structure of on-line links is transferred to the links-based indexes of Google which, consequently, favour the inclusion of established sites that enjoy a greater popularity expressed in the form of links.

Google's relevance criteria for selecting, ordering and omitting websites

According to information from the company (Google Corporate Info, n.d.), the search engine operates with 200 criteria that are weighted using a complex relevance algorithm. Some of them, such as the aforementioned key word density, may affect positively the representativeness of the sample. However, other publicly available criteria suggest that Google implements a hierarchical model of selection and ordering, which tends to favour dominant sites:

PageRank[3]

PageRank covers more than 500 million variables and 2 billion terms (Google Información Corporativa, n.d.) and is one of the most important calculations for determining the position of a web page in the search results. It is a numeric value (from 0 to 10) which Google assigns to each website it has indexed based on the quantity and quality of the links it receives. Although the details are unknown, the *PageRank* of a web page is determined by the number of links it receives, but also by the importance of the pages from which the link comes. Each link is rated more or less according to the *PageRank* of the source from which it comes (Google Corporate Info, n.d).

The links-based logic of selecting and ordering the possible sources also suffers from the skewing derived from the fact that the Internet is a scale-free network that obeys a power-law, just as occurs in the indexing process. One limitation is that heavily linked pages are more likely to be included in the universe than pages that receive few links, even if the information they contain is different, original and relevant to the query. Similarly, if less linked sources do enter the universe, they have less chance of achieving a good position compared to established pages. Moreover, links from consolidated sites with a high *PageRank* are ranked higher than those of new, minor and independent sites and those that are not part of the dominant circuits, which means that the votes of all participants do not carry equal weight. The dominant status of mainstream sites reinforces itself as it is easier for them to obtain better and more numerous links, because they are better known, because they maintain good relations with other dominant actors and because they can carry out on-line marketing campaigns for more links (Gerhards & Schäfer, 2010; Mager, 2009).

The result is that search engines like Google favour large, established sites of an institutional, governmental or commercial nature that are very well connected (Elmer, 2006; Introna & Nissenbaum, 2000), including in crucial areas such as health (Mager, 2009; Seale, 2005) or politics (Hindman et al, 2003).

Finally, Google does not distinguish between positive and negative links; as a result, the total number is not indicative of their informational value, or even of their importance in the field. Here we find at work the principle that any publicity is good publicity. This has led to some extreme cases that reflect the problems associated with indiscriminately adding up links to calculate the relevance of websites. For example, *The New York Times* (Segal, 2010) reported a series of malpractices committed by an on-line sales company. This company resorted to selling defective products and threatening to murder and sexually assault dissatisfied customers with the aim of generating negative on-line publicity that would improve the positioning of its website in searches. In this way, the company's owner managed to have his website linked to various consumer defence websites with a good *PageRank* on numerous occasions. In spite of the fact that the context of the link is fierce criticism of the company and condemnation of its illegal practices, the company succeeded in improving the positioning of its website in Google results and increasing sales.

Website loading time speed, age, size and updating frequency

Google also takes into account these factors, and by doing so, it offers a service as useful as it is discriminatory, as it is the sites with the largest economic and human resources that have the capacity to benefit from website optimisation. Moreover, although the relevance algorithm used by Google is kept secret, these criteria for application and others that are available in the public domain make the operations of the search engine vulnerable to *SEO* techniques. SEO stands for *Search Engine Optimisation* and *Search Engine Optimisers*. It refers to experts in obtaining a good position in the search engines for the sites of the companies and institutions that hire them. As a result, websites which can afford SEO have an optimisation tool that increases their chances of appearing in the results sample and appearing higher up than those that cannot afford it.

Summary of the logic on which Google operates

As mentioned above, it is not possible to know fully how Google functions. However, the criteria that have been identified and are in the public domain indicate that searches tend to benefit the websites of large, established commercial and professional institutions with economic and human resources, at the expense of new, minor, alternative, independent websites without much capital, although these also have some chance of being able to compete. From the perspective of the scientific validity of Google universes the following general observation can be made: As long as a random process to select the web pages is not possible, the representativeness and relevance of any sample provided by Google will be skewed.

Conclusion

A methodological procedure has been described which has been developed to find the materials included in universes of communications studies in the Google search engine. The same model is applicable to any other Internet universe. A description has been offered of how to ensure that the universe that Google retrieves has the greatest validity possible, adapting it to the criteria of relevance and uniqueness used in the research. A logical model of search structures has been presented which allows a systematic procedure to be followed to find the different sub-universes without producing repetitions or omissions. Google's search engine was chosen to test its efficiency, completeness, and representativeness. An analysis has been given of the limitations and problems posed by Google in providing the universes required for the research of the materials existing on the Internet. It has been noted that one of the most important inconveniences is that some publicly available criteria used by Google to select, order and omit websites generally favour dominant websites.

It is therefore concluded that the enormous progress represented by access to corpuses of sources that are so abundant and so fast and easy to use justifies the use of Internet search engines by researchers to obtain their corpuses. But it should be noted that search engines do not yet allow researchers to obtain complete universes or samples that are representative of the materials existing on-line. The skewing of samples that has been detected undermines, to varying degrees depending on the case, the reliability of all research involving the use of search engines.

This analysis of Google universes inevitably points to other more general conclusions. These relate to the controls that limit access to information, and to the possibility that search engines may sell, transfer or lose possession of information that could be used by agencies, governments and companies to limit the exercise of individual and collective rights and freedoms.

Joan Pedro is a PhD candidate in the program of Communication, Social Change, and Development, Universidad Complutense de Madrid (UCM), Spain. He is also a member of the interdepartmental research group 'Identidades Sociales y Comunicación' (Social Identities and Communication), UCM and holds a Master's Degree in Communication Theory and Research, from UCM.
www.ucm.es/info/secom, joan.pedro@pdi.ucm.es

Notes

[1] The Spanish version of Google has been used.
[2] Cf. Altman &Tennenholtz (2005), Brin & Page (1998), Brin et al (1998).

References

Altman, A., & Tennenholtz, M. (2005). Ranking systems: the PageRank axioms. *Proceedings of the 6th ACM conference on Electronic commerce*, Vancouver, BC, Canada, 1-8. Retrieved from http://stanford.edu/~epsalon/pagerank.pdf

Aubuchon, V. (2009). Google ranking factors - SEO checklist. Retrieved from http://www.vaughns-1-pagers.com/internet/google-ranking-factors.htm

Ayuda para Webmasters de Google. (s.d).Cómo calcula Google el número de resultados? Retrieved from http://www.google.com/support/webmasters/bin/answer.py?hl=es&answer=70920

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509-12. Retrieved from http://www.barabasilab.com/pubs/CCNR-ALB_Publications/199910-15_Science-Emergence/199910-15_Science-Emergence.pdf

Barfourosh, A., Anderson, M. L., Nezhad, H. R. M., & Perlis, D. (2002). Information retrieval on the World Wide Web and active logic: a survey and problem definition. *Technical Report, CS-TR-4291*. College Park, MD: University of Maryland, Computer Science Department. Retrieved from http://www.lib.umd.edu/drum/bitstream/1903/1153/1/CS-TR-4291.pdf

Benkler, Y. (2006). *The wealth of networks. How social production transforms markets and freedom*. London and New Haven: Yale University Press.

Bergman, M. K. (2001, August). White paper: The deep web: Surfacing hidden value. *Journal of Electronic Publishing, 7*(1). Retrieved from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the seventh international conference on World Wide Web*, Brisbane, Australia, *7*, 107-117. Retrieved from http://infolab.stanford.edu/~backrub/google.html

Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*, 29 January. Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Edelman, B., & Zittrain, J. (2002, October 26). Localized google search result exclusions. Statement of issues and call for data. *Berkman Center for Internet & Society*. Retrieved from http://cyber.law.harvard.edu/filtering/google/

Elmer, G. (2006). Re-tooling the network. Parsing the links and codes of the web world. *Convergence*, *12*(1), 9-19.

Fleischner, M. (2009). *SEO made simple: Strategies for dominating the world's largest search engine*. USA: Lightning Press.

Gerhards, J., & Schäfer, M. S. (2010). Is the internet a better public sphere? Comparing old and new media in the USA and Germany. *New Media & Society*, *12*(1), 143-60.

Google (2010). Search engine optimization starter guide. Retrieved from http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf

Google Corporate Info (n.d.). Technology overview. Retrieved from http://www.google.com/intl/en/corporate/tech.html

Google Centro para Webmasters (n.d.). Directrices para webmasters. Retrieved from http://www.google.com/support/webmasters/bin/answer.py?answer=35769

Google Información Corporativa (n.d.). Visión general de la tecnología. Retrieved from http://www.google.es/intl/es/corporate/tech.html

Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. *International World Wide Web Conference, Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan*, New York, NY: ACM. 902–903. Retrieved from http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf

He, B., Patel, M., Zhang, Z., & Chang, K. C. C. (2007). Accessing the deep web: A survey. *Communications of the ACM*, *50*(5), 95–101. Retrieved from http://brightplanet.com/images/uploads/Accessing%20the%20Deep%20Web%20-%20A%20Survey.pdf

Hindman, M., Tsioutsiouliklis, K., & Johnson, J. J. (2003). Googlearchy: How a few heavily-linked sites dominate politics on the web. *Annual Meeting of the Midwest Political Science Association*, Chicago, IL. Retrieved from http://www.cs.princeton.edu/~kt/mpsa03.pdf

Huberman, B., & Adamic, L. (1999). Growth dynamics of the World Wide Web. *Nature*, no. 401, p. 131.

Introna, L., & Nissenbaum, H. (2000). The public good vision of the Internet and the politics of search engines. In R. Rogers (Ed.), *Preferred placement: Knowledge politics on the Web* (pp. 25–47). Maastricht: Jan van Eyck Akademy.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the web. *Nature*, *400*, 107-109. Retrieved from http://www.cse.ust.hk/zsearch/qualify/DistributedSearch/acessibility%20of%20information%20on%20the%20web.pdf

Madhavan, J., Ko, D., Kot, L., Ganapathy, A., Rasmussen, A., & Halevy, A. (2008, August). Google's deep web crawl. *PVLDB*, 23-28. Retrieved from http://cseweb.ucsd.edu/groups/sysnet/miscpapers/p1241-madhavan.pdf

Mager, A. (2009). Health information mediated health: sociotechnical practices of providing and using online health information. *New Media & Society*, *11*(7), 1123-42.

Martín Serrano, M. (1974). Nuevos métodos para la investigación de la estructura y la dinámica de la enculturización. *REIS*, *37*, 23-83.

Martín Serrano, M. (1977; 2008). *La mediación social*. Madrid: Akal.

Martín Serrano, M. (1978a). *Métodos actuales de la investigación social*. Madrid: Akal.

Martín Serrano, M. (1978b). Un método lógico para analizar los significados. *REIS*, *2*, 21-51.

Martín Serrano, M. (1986; 2004). *La producción social de comunicación*. Madrid: Alianza.

Mediaciones sociales. (2007). *Número monográfico*, segundo semestre. Retrieved from http://www.ucm.es/info/mediars/MediacioneS1/Indice/indice.html

Pennock, D.M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, *99*(8), 5207-5211. Retrieved from http://www.modelingtheweb.com/

Schwartz, B. (2008). The Google quality raters handbook. Retrieved from http://searchengineland.com/the-google-quality-raters-handbook-13575

Seale, C. (2005). New directions for critical Internet health studies: Representing cancer experience on the Web. *Sociology of Health & Illness*, *27*(4), 515–40.

Segal, D. (2010). A bully finds a pulpit on the web. Retrieved from http://www.nytimes.com/2010/11/28/business/28borker.html

Sherman, C., & Price, G. (2001). *The invisible web. Uncovering information sources search engines can't see*. Medford, NJ: Information Today Inc.

Smarty, A. (2009). Let's try to find all 200 parameters in Google algorithm. Retrieved from http://www.searchenginejournal.com/200-parameters-in-google-algorithm/15457/

Sweeny, E., Curran, K., & Xie, E. (2010). Automating information discovery within the invisible web. In J.T. Yao (Ed.), *Web-based support systems* (pp. 167-81). London: Springer-Verlag.

Wouters, P. (2006). On the visibility of information on the Web: an exploratory experimental approach. *Research Evaluation*, *15*(2), 107-15.