SMART WEB VISIBILITY OF ORGANIZATIONS

Augusto Klinger, José Valdeni de Lima and José Palazzo Moreira de Oliveira
Instituto de Informática, UFRGS, Av. Bento Gonçalves 9500, Porto Alegre, Brazil

Keywords: Web Visibility, Smart Web Visibility, Universities Ranking, Homonymous Problem.

Abstract: Smart Web Visibility, in this paper, is the study of measure techniques used to retrieval information about words, expressions or terms (for example, acronyms) on the web. The visibility has straight relation with presentation order of information retrieved. We consider the Smart Web Visibility as a subfield of the Webometrics, the same as Web Visibility is its subfield too. Our approach is based on different rankings from different search engines to evaluate the Smart Web Visibility by processing the homonymous problem before scoring. We begin with original results of search engines output, to emphasise after with methods to adding semantics to the queries. Finally, to demonstrate the viability of our ideas we employed acronyms of Brazilian universities to evaluate the smart visibility and compare with the actual situation of Brazil universities published by Webometrics Ranking. The main contribution of this work is a new way to evaluate the Web Visibility, named Smart Web Visibility, which shows how the universities are ranked by multiple search engines.

1 INTRODUCTION

Visibility on the Web is a quantitative measure of the visibility of a webpage on the network, determined by the ease that Web users find it. The study area of measures of visibility on the Web is named Web Visibility considering the quantitative aspects of construction and use of information on the Web viewed under bibliometric aspects (Björneborn and Ingwersen, 2004). The goal of Webometrics is to obtain information by measurements on the various aspects of the Web obtaining, for example, statistics data about popularity, clusters of websites and distribution of information. We named Smart Web Visibility the visibility provided by the different visions of the eyes of the Web, the search engines.

Web visibility measuring is important in several respects, principally to evaluate the level of advertising or measure the impact of a trademark, product or institution in the network. There are different ways of measuring the visibility of a website: number of hits, number of links that lead to the website, or position in the ranking of a search engine.

The measure most spread in the literature is the use of the number of web links pointing to a webpage, or inlinks, as a measure of visibility in the network (Aguillo, Granadino, Ortega and Prieto, 2006; Aguillo, Granadino and Ortega, 2006). This measure is also used to generate the ranking of universities by the Cybermetrics Lab, the Webometrics Ranking of World Universities. The criterion of number of inlinks is usually used in the algorithms that assemble the rankings of search engines, as the world famous PageRank (Page, Brin, Motwani and Winograd, 1999). Few studies use the number of accesses to a webpage as a measure of visibility (Aaltojärvi, Arminen, Auranen and Pasanen, 2008) and there isn't widespread works using search engine's results for the calculation of visibility on the web, although they are the main option for any web search. The development of an approach to evaluate visibility based on web search engines is presented as an interesting alternative to the existing evaluation web visibility methods. The measurements are relevant not only to the field of web marketing, but also for the elaboration of rankings in certain domains and evaluations of popularity of general purpose on the web.

2 RELATED WORK

A work of Aguillo (2006b) analyzed the presence of Brazilian universities on the web. According to the author, the developing countries of Latin America are making efforts to publish electronically the results of their researches and studies. The size of their web domains has grown, as well as its visibility on

Klinger A., Valdeni de Lima J. and Palazzo Moreira de Oliveira J..
SMART WEB VISIBILITY OF ORGANIZATIONS.
DOI: 10.5220/0003928906710676
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 671-676
ISBN: 978-989-8565-08-2

the web. The results show an increasingly strong presence of Brazilian universities on the web, but still far from developed countries. The study data were obtained from eight search engines. The indicators were the number of pages (size), number of inlinks (visibility) and number of visits (popularity). The São Paulo University (USP) has led the rankings according to three indicators. University of Campinas (UNICAMP) was second in size and visibility, the second in popularity is the Federal University of Rio de Janeiro (UFRJ).

Web visibility indicators were examined to assess how far the collaboration in science and technology publications are visible on the Web (Aguillo and Kretschmer, 2004). The study found that about 80% of the bibliography with multiple authors is visible through search engines. The studies by Kretschmer (2007) show that the structures of hyperlinks does not reflect the collaborative structures of bibliographic data, but web visibility indicators are different from hyperlinks and can be used successfully as indicators of web collaboration.

Barjak and Thelwall (2008) report the results of a study of the connection between the count of inlinks and the real significance of the websites of about 400 research groups in Europe. The analysis confirmed that the size of research groups and their presence on the Web are important to attract links, while the scientific production itself is not. The interpretation of data from search engines need to be further studied before to take conclusions about its usefulness as indicators, conclude the authors.

However data from search engines are widely used in several works with different purposes. A quick web search reveals many applications such as: domain evaluators (Dnscoop, 2009; Cubestat, 2008), which provide a dollar amount to the target site; trees of words, which are based on user queries of search engines (Viegas, n.d.).

The works of Espadas, Calero and Piatinni (2008) and Gori and Witten (2005) highlight important points of Web Visibility. The first deals with the problem that search engines do not make large parts of the Web visible, proposing a method for indexing sites. The second explains some heuristics adopted by search engines and expose their weakness by allowing the construction of artificial communities of sites that link to each other in order to improve their rankings. A possible solution to the problem is in the Semantic Web.

A previous work used notions of relevance and precision of metasearch engine rankings combined with a rankings fusion method to develop a calculation of Web Visibility (Klinger et. al., 2011). The results serve as an indication that the web search engines provide interesting data for the evaluation of visibility and point to future studies to apply and expand the formula in a general way.

3 METASEARCH ON THE WEB

This work aims to define a new way to evaluate visibility on the Web based on information collected by several search engines. As the volume of data on the Web is very large, and growing, no search engine can index the entire web. Additionally, only the best placed websites are displayed to users in efficiency concern. What we have as the result of a search for a particular term in any web search engine is a classification according to the search engine used, covering a portion of the web, classified according to their heuristics, techniques and proprietary algorithms. One way to increase the scope of coverage of the Web is the utilization of more search engines. This process of consulting several search engines is known as metasearch. As the search engines are the dominant points of access to webpages, the metasearch engine presents itself as an interesting tool for measuring visibility on the web, resulting in a quantitative data representing how accessible and how well regarded is the website in accordance with the web search engines. Using more search engines, tends to increase the diversity of criteria and range, generating a more reliable value of web visibility.

Considering the official website of an organization as the major milestone of its presence on the network, is expected that when a search engine is consulted for the organization name, or its acronym, the official website is between the first placed results. This means that the organization website, and therefore its name, has a good visibility under the search engine used. By applying a metasearch with any term (or list of terms) the results are different rankings, one for each search engine to which the metasearch system forwarded the query. Usually a single ranking is displayed to the user, making necessary a fusion technique for the various rankings. There are several methods to merge different classification functions, Rankings Fusion. However, as we haven't interest in see a ranking of all the webpages retrieved, we want only a value of visibility, there is no need to merge rankings. What matters for the evaluation of visibility on the Web are just the placements in the various search engines of the organization official website that we are calculating Smart Web Visibility. In the evaluation of web visibility, the metasearch engine is used as follows: the

name or acronym of the organization you want is consulted at *n* search engines; then the official website of the organization (previously known) is located in each of the *n* classifications, featuring a second moment of searching; finally the visibility of the organization is scored according to the official website placement in the search pages rankings.

The evaluation method scores the placement of webpages in search engines as the classic method Borda Count (Saari, 1985), used in voting processes (Black, 1976) and also widely applied to computational problems such as rankings fusion and metasearch (Aslam and Montague, 2001). For making use directly of the placement of the elements of a ranking as means of scoring, the method satisfies the needs of the proposed evaluation way of web visibility. Borda Count is, basically, an election method in which each elector ranks the candidates in order of preference. The winner is determined by points given to each candidate according to the position they are in the list of preference of each voter. The candidate with the highest score at the end of the count is the winner. To determine the score for each placement, we need to know the number of candidates. Thus, for *n* candidates, the first receive *n* points, the second *n-1*, and so on. Alternatively, we can assign one point for the first place, 1/2 for the second, 1/3 to third, and so on, giving emphasis on the first place. Table 1 below illustrates the first scoring method, for $n = 5$.

Table 1: Scoring by Borda Count.

Placement	Formula	Score
1º	n	5points
2º	(n-1)	4points
3º	(n-2)	3points
4º	(n-3)	2points
5º	(n-4)	1 point

As the elements that will be voted in the context of this work are webpages, we can't know all the candidates. The voters are search engines and they tend to rank a varying quantity of webpages. Also, the rankings do not necessarily contain the same webpages.

The solution for the problem is to use only the first places of each search engine, known as top-n. As the rankings will be used for evaluate visibility and it is known that users of search engines tend to focus only on the first places, truncating the results should not affect the results nor distort the value calculated. According to our previous studies, work only with top-10 of the search engines is ideal, because as we increase the value of *n*, the greater is the noise in the rankings, i.e., increases the number of unwanted sites (irrelevant) returned by the query. Thus, using the top-10, the first placed website receives ten points, the second nine, until the tenth that receive just one point.

3.1 Evaluating Web Visibility

For a given organization to which we want to measure the visibility on the web, the evaluation of web visibility helped by metasearch and based on the positions of the official website of the organization in the various rankings proceeds as follows: i) Identification of the organization official website; ii) Search by acronym (or name) of the organization in *n* search engines; iii) Search for the official website in each of the *n* rankings; iv) Sum of scores according to the position in each ranking.

We have implemented a prototype for experimental purposes, so we can efficiently produce rankings of institutions belonging to the same domain. In the prototype were included fourteen search engines. The input parameters are the target of the search and the website. The query is sent to fourteen search engines and the top-10 of each ranking are placed in a matrix where each column represents a search engine and each row represents a retrieved webpage. As among the top-10 of each search engine does not necessarily appear the same ten pages, a zero value is assigned to cells in the matrix that correspond to pages that did not appear in the top-10 search engine column. In the other cells of the matrix are assigned the placement of the websites according to the rankings. Of this matrix is utilized only the row corresponding to the organization's official website. For each non-zero value of the line are assigned and summed points, according to the Borda Count, reaching a maximum value of 140, in which case the official website returned in first place in all fourteen search engines. The search engines involved are: Brazilian versions of Alta Vista, Ask, Google and Yahoo, global versions of Alexa, All The Web, Alta Vista, Ask, AOL, Exalead, Google, Icerocket, Lycos and Yahoo. The choice of search engines will affect directly the results, so this is a very important step.

A problem that can occur with this way of measuring visibility is when there are other organizations using the same name or acronym. This reduces the visibility value calculated and characterizes the problem of homonyms, since the organizations will compete for positions in the same query in the search engines. A way to eliminate the problem is specifying the domain, adding semantics to query in the case with the query expansion including academic terms.

4 UNIVERSITIES RANKING

Universities characterize a homogeneous domain, where each institution has an acronym and a webpage that does not vary very much from certain standard, being an ideal study case for ranking organizations based on the vision of their official websites by search engines. The use of information from the Web to rank universities is nothing new. The QS World University Rankings, by QS Quacquarelli Symonds Limited, uses Scopus, which is a database (available in Web version) of abstracts and citations of scientific literature production, to measure the intensity of research through the documents recoverable by the platform. The Academic Ranking of World Universities (ARWU), by Shanghai Ranking Consultancy, uses data sources of the Web to define their classification criteria which involve number of publications, citations and awards received by the researchers. Another ranking of universities worldwide is the Webometrics Ranking of World Universities, by Cybermetrics Lab, which uses Webometrics and its sub-areas, particularly the Web Visibility. In its rankings, the Web Visibility represents 50% of the total score aggregate to the university, and this visibility is measured by Yahoo! Search, taking into account the total number of unique inlinks that each official university's website receives.

There are several rankings classifying higher education institutions worldwide on the web, as can be seen in the work of the Nordic research council Nordforsk (2011). The analysis of the presence of the universities by means of cybermetric indicators shows up as an important tool for evaluations and comparisons, being increasingly more relevant. A good placement within a ranking can attract more high-level researchers, students and investment for the university. Universities have become aware of the importance of their presence on the web. A way to maximize the visibility of an institution is to maintain a digital repository that represents the scientific output of the institution (Swan and Carr, 2008).

The following section is the applying of the formula developed by this work to the specific domain of the Brazilian universities.

4.1 Brazilian Universities Ranking

Thirty Brazilian universities were submitted to the Smart Web Visibility evaluation. The universities were chosen based on the set of Brazilian universities of the Webometrics Ranking of World Universities for future comparisons. The acronym of each university was used as a query parameter in the prototype developed, along with their respective official websites previously identified, revealing the visibility of the acronym linked to the university in the web. Table 2 contains the top-10 universities sampled in the experiment, 140 being the maximum visibility value, corresponding to fourteen first places.

In the ranking of table 2, containing the fifteen best placed universities among the thirty submitted to the Web Visibility evaluation, there is the Pontifical Catholic University of São Paulo (PUC-SP) as leader and the one with maximum points. After is the Pontifical Catholic University of Rio de Janeiro(PUC-Rio) showing that both acronyms have a great power of discrimination and excellent visibility in the eyes of search engines. The Federal University of São Paulo (UNIFESP) completes the top-3, followed by five universities tied in fourth place, two tied in ninth place, UNICAMP in eleventh, three universities again tied in twelfth and UFPR completing the top-15.

Table 2: Top-15 sampled universities.

Rank	University	Score
1º	PUC-SP	140
2º	PUC-Rio	138
3º	UNIFESP	137
4º	UFRGS	135
5º	PUCRS	135
6º	UFRN	135
7º	UFSM	135
8º	UERJ	135
9º	UFRJ	134
10º	UFSCAR	134
11º	UNICAMP	133
12º	UFPE	131
13º	UFPB	131
14º	UNISINOS	131
15º	UFPR	130

It is interesting to compare the results with the ranking of the Brazilian institutions on the Webometrics Ranking of World Universities. Of the top-3 ranking, USP and UNICAMP (1st and 2nd, respectively), did not make the top-10 in this experiment. Considering that they are two major universities in Brazil, and also that among the webpages returned by metasearch featured many that are unrelated to universities, a new form of querying was experienced. The homonymy problem was evident in this first experiment. Some universities were affected, as the Federal University of Ceará (UFC) whose acronym also belongs to an organization most famous, the Ultimate Fighting Championship. In this first ranking the UFC was in the thirty position. In a

new experiment, the acronyms of the same thirty universities were resubmitted to the query adding the word 'university'. This process is known as Query Expansion, and serves to define more precisely what we are looking for, in this case universities, avoiding the homonymy problem. The new ranking is shown at table 3.

Table 3: Top-15 sampled universities with query extension.

Rank	University	Score
1º	UNICAMP	140
2º	UEM	140
3º	UFV	140
4º	USP	136
5º	UFRGS	136
6º	UNB	136
7º	PUC-Rio	136
8º	UFPE	136
9º	UFF	136
10º	UFRN	136
11º	PUC-SP	136
12º	UFPB	136
13º	UNESP	135
14º	UFSC	135
15º	UFC	135

Now, in the first place was a tie of three universities, which were not ranked in the top-10 before, and they obtained maximum score. UNICAMP, together with the State University of Maringá (UEM) and the Federal University of Viçosa (UFV) had a score of 133, 77 and 83 in the previous experiment, respectively, obtaining the maximum score in the version with the expanded query. Next, nine universities were tied with 136 points, among them the PUC-SP and PUC-Rio, showing that the technique has increased the noise level in the rankings by search engines, which already had a great precision for these acronyms, but not impaired, as they continue well placed. Also in the ranking of table 3, there is a good placement of the USP, which rose from 63 to 136 points, entering the top-10.

In general, the word 'university' with the acronym in the query raised the score of all universities sampled. In the new version of the experiment, the university with less scoring scored 90 points, in contrast to 48 points in the first experiment. The total score of all the thirty universities, using the query only by the acronym were summed 3466 points, using the expanded query the sum reached 3910 points. The graphic of figure 1 shows the increasing in scores for some of the most prestigious universities of Brazil.

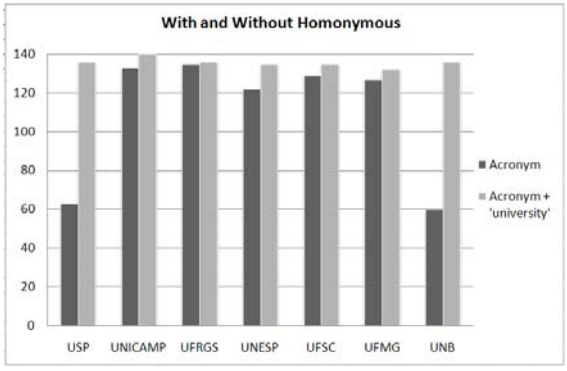

Figure 1: Scoring with and without the homonymous problem.

The homonymous problem of universities like UFC was avoided by Query Expansion and allowed the university to rise from the thirty position to the tenth fifth position. Of course, if other organization linked to universities has the same acronym the problem would not be solved by simple adding the word ‘university’ in the query. A more sophisticated query expansion would be necessary to include extra semantics.

5 CONCLUSIONS

As seen, the visibility of an organization on the Web can be measured in several ways. The most common form is the count of the unique external inlinks, as used by Webometrics Ranking. Another way to measure the presence on the network can be represented by the number of webpages recovered from search engines or articles indexed. The universities rankings of the Internet use mostly bibliometric indicators, especially citation.

The main contribution of this work is a new way to evaluate visibility on the web, an indicator based on data from search engines. When performing a search on any search engine, users tend to look only at the first results. That is, a website well placed in a search engine has, in other words, a good visibility in such search engine. This idea was explored: the placement of the website in various search engines, rather than the number of pages on the domain of the institution, number of documents, citations or links. It was presented a way of evaluate visibility on the Web through search engines, taking into account the placements of the webpage linked to the search argument in several rankings. The evaluation way presented was named Smart Web Visibility and shows how well a particular entity is perceived by web search engines.

Through a study case in order to rank universities by Smart Web Visibility, we observed an interesting application of the evaluation proposed, showing a current scenario that is the subject of several researches. Applying metasearch on the universities'acronyms, two rankings were developed: one showing the visibility of webpages of institutions when a search made with only the acronym, and another using a query expansion technique to better describe the domain, increasing the scoring of the universities sampled in the experiments and avoiding the homonymous problem.

The Smart Web Visibility has applicability in any field, not only universities, but for the generation of rankings is important that the domain is homogeneous. Future studies should seek a way to demonstrate the ampleness of the method.

5.1 Future Work

As mentioned above, efforts are still required to prove the application of Smart Web Visibility evaluation generically, allowing us to develop rankings in other domains. Furthermore, the work identified the possibility of some future studies like the study of other parameters that can be extracted for the evaluation of web visibility, the study of tiebreakers for visibility rankings, and the study of a distribution of different weights to each search engine according to some criterion to be studied too. Future works will be concerned about two main topics. One of them is to add more semantics to the description of the domain, perhaps by ontologies, making possible to navigate through the domain levels. The other main topic is about extracting time and spatial data with the metasearch, aiming to discover where and when the visibility of the target was better or worst. In the near future, rankings with more universities, including universities outside of Brazil, should be developed.

ACKNOWLEDGEMENTS

This work has been partially supported by CNPq and by CAPES, Brazil.

REFERENCES

Aaltojärvi, I., Arminen, I., Auranen, O. and Pasanen, H-M.(2008). Scientific Productivity, Web Visibility and Citation Patterns in Sixteen Nordic Sociology Departments. *ActaSociologica*, 51(1), 5-22.

Aguillo, I. F.,Granadino, B., Ortega, J. L. and Prieto J. A. (2006a). Scientific Research Activity and Communication Measured with Cybermetrics Indicators. *Journal of the American Society for Information Science and Technology*, 57, 1296-1302.

Aguillo, I. F., Granadino, B., Ortega, J. L. (2006b). Brazil Academic Webuniverse Revisited: A Cybermetric Analysis. In *Proceedings… International Workshop on Webometrics, In-formetrics and Scientometrics& Seventh COLLNET Meeting.* Nancy, France.

Aguillo, U. F. and Kretschmer, H. (2004).Visibility of Collaboration on the Web. *Scientometrics*, 61, 405-426.

Aslam, J. A. and Montague, M. (2001).Models for Metasearch.In *Proceedings… ACM SIGIR Conference on Research and Development in Information Retrieval*.SIGIR'01. ACM, New York, NY, 276-284.

Barjak, F. and Thelwall, M. (2008). A Statistical Analysis of The Web Presences of European Life Sciences Research Teams. *Journal of the American Society for Information Science and Technology*, 59, 628-643.

Björneborn, L. and Ingwersen, P. (2004).Toward a Basic Frame-work for Webometrics. *Journal of the American Society for Information Science and Technology*, 55, 1216-1227.

Black, D. (1976). Partial Justification of the Borda Count. *Public Choice*, 28(1), 1-15.

Cubestat. (2008).*Cubestat*: The Free Website Value Calculator. Retrieved in November 21, 2011, from http://www.cubestat.com

Dnscoop. (2009). *Domain and SiteValueTool*.Retrieved in November 21, 2011, from http://www.dnscoop.com

Espadas, J.,Calero, C. andPiattini, M. (2008). Web Site Visibility Evaluation. *Journal of the American Society for Information Science and Technology*, 59, 1727-1742.

Gori, M. and Witten, I. (2005). The Bubble of Web Visibility. *Commun...* ACM, 48, 115-117.

Kretschmer, H., Kretschmer, U., Kretschmer, T. (2007). Reflection of Co-Authorship Networks in the Web: Web Hyperlinks Versus Web Visibility Rates. *Scientometrics*, 70, 519-540.

Nordforsk.(2011). *Comparing Research at Nordic Universities using Bibliometric Indicators.* NORIA-net. Retrieved in August29, 2011, from http://www.nordforsk.org/files/rapp.bib.2011.pub_21.5.11.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking*: Bringing Order to the Web. Technical Report. Stanford InfoLab.

Saari, D. G. (1985). *The Optimal Ranking Method is the Borda Count*. Discussion Papers, (638). Northwestern University.

Swan, A. and Carr, L. (2008) Institutions, their Repositories and the Web. *Serials Review*, 34, 31-35.

Viegas, F. B. (n.d.). *Word Tree*. Retrieved in November 13, 2011, from http://fernandaviegas.com/wordtree.htm