Social-network tools for the assessment of the university web performance

José Luis Ortega

R&D Analysis, Vice-presidency for Science and Technology, CSIC, Madrid, Spain

Isidro F. Aguillo

Cybermetrics Lab, CCHS-CSIC, Madrid, Spain

Abstract

This contribution intends to introduce Webometrics as an emerging discipline focused on the understanding and assessment of the academic information flows on the Web. It describes the principal web-based techniques and tools used to evaluate the performance of higher education websites and to explain how these information networks are created and modelled. This chapter starts with an introduction to the Webometrics, where we present its origins and evolution, its theoretical framework and its relationship with other web disciplines. Next, we describe the principal indicators and measures used to quantify the development of several web units (web domains, sites, pages, etc.). We mainly stress the properties of the social-network measures in order to describe the visibility of a web site and to characterize the structure of a web space. We continue with a description of the main developments such as the Ranking of World universities on the Web and visualizations of web regions. Finally we finish with a discussion about the implications of this discipline in the improving of the web performance and visibility of the university institutions on the Web, and its impact in the development of the higher education web-based policies according to open access and e-learning initiatives.

Webometrics: a discipline devotes to quantify the web performance

Webometrics is a young discipline born around mid-1990s with the seminal work of Almind and Ingwersen (1997) and the creation of the first specialized e-journal on webometric studies, *Cybermetrics*. It emerged in a moment in which the Web has settled down in the academic world and it starts becoming a new and powerful way to communicate scientific results. As scientometrics is focused on the assessment of print-based communication processes (papers, patents, citations, etc.), Webometrics is targeting web-based communication units such as web domains, pages and hyperlinks as a way to understand new scientific activities including those unrelated to the print world. Thus, Webometrics is defined as "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (Björneborn and Ingwersen 2004). The first works tried to apply the research evaluation to the Web, looking for relationships between web production and visibility with scientific activity and impact. Thus, a strong correlation was found between the web pages / link ratio and the scores of the research assessment exercises in the United Kingdom and Australia (Thelwall 2001; Smith and Thelwall 2002). Significant relationship were also found between links and journal citations (Vaughan and Thelwall 2003), and between web-based university rankings and rankings built on bibliometric data (Aguillo et al. 2006). These papers gave the discipline a great soundness, allowing to understand non-formal scientific communication process on the Web and their relationships with other scientific outputs (papers, books, patents, etc.). The growth of the Web and the incorporation of these traditional formats to the Web prompted the appearance of studies about formal scientific communication on the Web such as impact of e-journals (Harter and Ford 2000), scientific repositories (Antelman 2004) and web-based citation indexes (Google Scholar, Scirus, etc.) (Bar-Ilan 2008).

One of the most concerning issues related to the Web studies was about the reliability of the data sources used to develop quantitative analysis and the meaning of the results obtained. Basically, search engines and crawler data were used to carry out web researches. The appearance of AltaVista in 1995 and its search operators suggested that the search engines may be used as a web citation index (Rodriguez-Gairín 1997). After that several studies showed that the search engines were unstable along short time periods (Rousseau 1998); their operators were weak and their databases frequently outdated (Sullivan 2003). Other contributions detected linguistic biases in non-Latin languages (Bar-Ilan 2005) and a low overlap between search engines (Lawrence and Giles 1998). This situation favoured the use of web crawlers, customized for the harvesting process and direct extraction of exhaustive information about a website. On the contrary, they consume a lot of time and technical resources, as well as they evidenced difficulties to extract and follow links from non-textual formats (Chakrabarti 2002). After the search engines war in 2003, the largest engines improved their stability and their search operators reported more consistent results (Bar-Ilan 2009). Thus, search engine data are used to develop broad scope studies because allow obtaining huge amount of quantitative data at the level of countries and domains, while the crawler data are suitable to carry out micro studies on web sites and link content.

However, Webometrics have to face the volatile nature of the Web in which the contents appear, change and vanish in a short time period (Ortega et al. 2006) and where a rate of web page disappearance of 0.25% to 0.50% per week evidences a highly changing world (Fetterly et al. 2003). This instability attracted the attention of many studies that try to understand such phenomena, investigating the ephemeral existence of incoming links in e-journals (Harter and Kim 1996), web citations in scientific repositories (Lawrence et al. 2001) and web content decay (Payne and Thelwall 2008). These studies can be defined as Web demography because they observe the web as a population of contents that born, growth and dead along the time. In this way there are studies that calculate the age of the Web (Ortega et al. 2009) the ratio of change of web pages (Cho and García-Molina 2000) or the death of web pages (Koehler 2004).

The analysis of the information usage of web sites have attracted early attention from business and commercial web sites interested in gathering and processing information about the behaviour of their customers (Gomory et al. 1999), as an extension of the data mining techniques applied to their client databases. This field has not been exploited in depth by the scholars mainly due to the difficulty of obtaining the log data and comparing similar patterns of different web log sources. Several works focused in analysing the search skill and attitudes of the principal search engines' users such as AltaVista (Silverstein et al. 1998), Excite and Alltheweb (Jansen et al. 2005) and Yahoo! (Teevan et al. 2006), while others targeted methodological problems like definition of web sessions and the advantages of using them instead of the number of hits. Data mining was used for the identification of web sessions, to estimate their duration and their length in clicks (Pitkow 1997), to classify content according to the pages requested by their visitors (Wang and Zaïane 2002) and to show navigational differences between different point of access (Ortega and Aguillo 2010).

Recently, a new way of understanding the web services and relationships has emerged, the so-called Web 2.0. In this new paradigm, the Web is becoming a way of collaborative creation of contents in which the freely active web surfers contribute personal experiences and own contents. This favours the emergence of web sites which principal characteristics are the person interaction in the contents design and the participative relationships between those users. This new environment gives the opportunity to study how the online environments affect to the social relationships (Lenhart and Madden 2007), what structural differences exist with other large-scale networks (Kumar et al. 2006) or what contents characterize these networks (Thelwall 2008).

Structural indicators: Social network measures as web indicators

The Web is essentially a huge network of interconnected webpages through hyperlinks which allow us to navigate sites and domains around the world looking for relevant information. The interconnection degree of a web site is key to be reached by potential users as the more incoming links (visibility) a web site receives the larger is the likelihood to achieve more visitors (popularity). Furthermore, according to the link popularity of the website that connects ours the probability of being located and visited also increase. Hence, not only is important to be connected but also to know who is linking us. This structural characteristic of the Web is essential to understand the position and successful of a web site. So, the use of the Social Network Analysis (SNA) has been crucial to go in depth in the assessment of web sites.

The Web as a graph

When Tim Berners-Lee named “World Wide Web” to the hypertextual information system developed in the CERN (*Conseil Européen pour la Recherche Nucléaire)* he sensed that that system would be a complex web-shape network in which each html document will be a node connected to the whole repository through hyperlinks, but what would be the shape of that network? And what importance would be the shape of that system? The large size of the Web in number of pages and links and the easiness of harvesting this information through web crawlers attracted the attention of many scientists that want to empirically observe if the Web followed a random network shape. They were surprised when observed that there was not a constant parameter or scale in the degree distribution such as random networks but it followed a potential distribution (power law), in which there are a small number of highly connected nodes while the remaining ones have barely a few links (Barabasi and Albert 1999). These scale-free networks also show a high clustering coefficient and a short average path length as the small world networks, which means that the Web is a decentralized environment where there are a high density of links and where highly connected nodes (hubs) supporting that density emerge. Barabasi and Albert (1999) suggested that the formation and evolution of the scale-free networks is due to the “preferential attachment” phenomenon, which states that the best connected nodes are more likely to obtain new links than the less connected ones. This phenomenon provokes skewed distributions and the emergence of large hubs that bring together the network. Other factors that affect the emergence of scale-free networks such as competition and fitness (Bianconi and Barabasi 2001), optimization (Valverde et al. 2002) or uniform attachment (Pennock et al. 2002) were found. However, these factors do not take into account web contents and other sociocultural phenomena that would explain the dawning of search engines or web 2.0

sites. Thelwall (2002) found that there is a geographical pattern in the link relationship between British universities web sites and Ortega et al. (2008) observed that the language is a strong variable for explaining interlinking among university web domains.

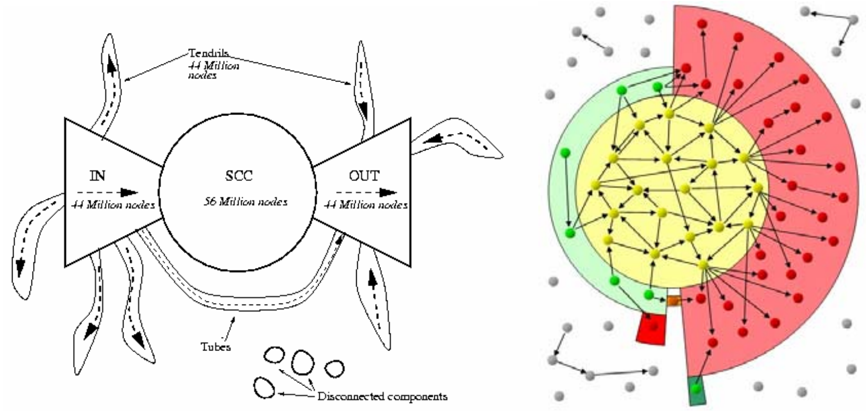

Figure 1. Topological proposals about the form of the Web. "Bow-tie" model (Broder et al. 2000) and "corona" model (Bjornerborn 2004).

Nevertheless, the Web is a directed network in which the orientation of the links does not need to be reciprocal. Thus, a web site may be linked by a lot of web pages but it does not link to any other. Taking this into account, several models were proposed in order to study the web topology. The "bow-tie" model (Broder et al. 2000) localizes the web pages in four regions according to their link relationship with the other ones: the SCC or Strong Connected Component is the zone where all the nodes are connected among themselves; IN component show nodes that link to SCC but they are not reached from SCC; in the OUT component, the nodes are linked from SCC but they do not tie to SCC; and the TENDRILS are nodes that link to other nodes outside the sample. This model allowed characterizing a web space (geographical, thematic, etc.) studying its components' size. Hence, a very large SCC component shows a highly compact environment while a big OUT component is a sign of dependency with other web space. The "corona" model (Bjornerborn 2004) is a variation of the first one in which the IN and OUT zones are directly related.

Social network indicators

The interconnected nature of the Web forces to the web researchers to adopt structural indicators in order to measure the web activity of a web site, domain or space. These structural indicators allow us to define properties of the analysis units and compare their performance into the Web. SNA techniques have helped the development of web indicators that measure the structural relationship of a web site with its surroundings or to study the main characteristics of a web space. Next, we detail the most important social network indicators used in Webometrics:

Individual indicators

These indicators are focused in the situation of a node in the network; they describe the importance and meaning of a vertex in the context of whole the network.

- **Centrality Degree**: It measures the number of lines incident with a node, that is, total number of links that a web site, domain or space receives. This can be normalized (nDegree) by the total number of nodes in the network. Since the Web is a directed network, we can only count the incoming links (InDegree) or the outgoing links (OutDegree). The incoming links are sign of visibility because they generate traffic and visits to a web site, raising its popularity. Furthermore, the in-links are considered as a prestige indicator because they can be interpreted as an authoritative citation. On the other hand, the outgoing links show the mediator property of a website which directs the navigation to new web sites, domains or spaces. Following the Kleinberg’s (1999) nomenclature, the very in-linked sites are defined as Authorities, while the greatest out-linkers web sites are called Hubs. When a network is built from aggregated data, i. e. network of web domains, countries, regions, etc., each tie between two nodes represent the total amount of link from all the web sites of a domain, country or region to another one. In this weighted networks the centrality degree is calculated as the sum of the weight of each tie connected to a node. It was used by Kretschmer and Aguillo (2005) to highlight the scientist presence on the Web and gender differences, while Ortega et al. (2008) used it to rank the most out and inlinked European universities.

- **Betweenness Centrality**: It measures the intermediation degree of a node to keep the network connected, that is, the capacity of one node to connect only those nodes that are not directly connected to each other. In weighted network the Dijkstra's algorithm help to select the shortest path and hence to calculate the betweenness centrality according to that path. From a webometric point of view, this measure allows us to detect hubs or gateways that connect different web sub-networks. It was used by Björneborn (2004) to observe small world phenomena in the British academic web and by Ortega et al. (2008) to detect European web universities that mediate between their local sub-network and the European one.
- **Closeness Centrality**: It is an indicator that measures the average distance in number of clicks of a node with every node in the network. It is good indicator to study infection processes and information flows, because this centrality is based in the proximity of a web site with the rest. A high closeness shows a high reachability of a website during a navigational process. Dijkstra's algorithm is also used in weighted networks in order to calculate the closeness centrality. This index was used by Chen et al. (2006) to detect prominent members in a mailing list.
- **Eigenvector Centrality:** It indicates the relevance of a node according to the importance of other nodes that link it. This is a recursive indicator that transmits the value of a node to their acquaintances. It is a prestige index that does not only value the quantity of partners but the importance of those. An adaptation of this indicator was the popular PageRank (Brin and Page 1998) developed for positioning the most valuable pages in the top of the query results of Google.

Network indicators

These indicators measure the main characteristics of the network in all, describe the relationships of whole the members between them. They allow us to compare and to know how a network is structured.

- **P-Cliques**: a *p*-clique is a sub-network where every node is directly connected with the other ones. It shows groups with a high density and it is a way to detect underlying sub-networks. The value of *p* corresponds with the number of nodes that constitute a clique. It was used by Cothey, Aguillo and Arroyo (2006) to uncover web site structures clustering web pages, while Ortega et al. (2008) use it to identify national and regional groups in the European web space.
- **K-Cores**: is a sub-network in which each node has *k* degree in that sub-network. Unlike the *p*-cliques the *k*-cores allow us to detect groups with a strong link density. In the scale-free networks such as the Web, the core with the highest degree is the central nucleus of the network, detecting the set of nodes where the network rests on. Ortega and Aguillo (2009) used this measure to detect what universities make up the centre of the world academic network.
- **Distance**: is the number of steps in the shortest path that connect two nodes, the average among all the shortest paths in the network is the average distance. A short mean average distance is a good indicator of network density. Broder et al. (2000) applied this measure to show the density of the Web, finding an average distance of 16 clicks.
- **Diameter**: is the number of steps in the longest path. Just like the distance allows us to measure the cohesion of a network because it shows the largest distance that a node has to cover to reach the most distant node. Diameter was also used by Broder et al. (2000) to measure the thickness of the Web, while Björneborn (2004) applied it to detect “small-worlds” properties on the Web.
- **Global Clustering Coefficient:** it is a measure that shows the density or cohesion of the Web. It shows the proportion of nodes that tend to group together. Mathematically, it is the proportion of

closed triads by open triads, a triad being a group of three linked nodes. This measure is important to detect "small world" phenomena on the Web.

Visualising the Web

Several approaches have been used to present a visual picture of the Web that allows understanding how their elements are related between them and what are the principal structural characteristics of the Web. Next we summarize some of the most important.

Co-link Analysis

The first attempt was to represent web sites relationships through co-links. This technique studies the number of co-occurrences of linked web pages, sites or domains on a certain link corpus. Co-link Analysis assumes that if two web units appeared together then they are somehow related between them. To apply this technique, a co-occurrence matrix has to be built from search engines or crawler data. If we are collecting data from a search engine, then we recommend asymmetrical matrices be used as the links analysed belong to the study population, while a symmetric matrix counts the links from all the web sites indexed in the search engine database which introduces noise and biases. Then, a proximity measure (Salton's cosine, Spearman's rank correlation coefficient, etc.) is applied to transform the data into a distance matrix. Finally, we use a statistical model to project these distances in two dimensions, usually Principal Component Analysis (PCA) –a method to reduce several correlated variables to a few of components- or Muldimensional Scaling (MDS), which builds a point map according to the distances between the objects in an iterative process.

This technique is really more of a location method than a visualization one, because their proximities are presented as (x,y) coordinates and then may be plotted together with other visual elements such as links, size, shape, etc (Figure 2). Co-link is mainly used to detect content relationships between web units. Hence, Larson (1996) observed thematic clusters in web pages about geographic information systems, while Vaughan (2006) detects in the Canadian university web sets of universities by their cultural and linguistic relationships.

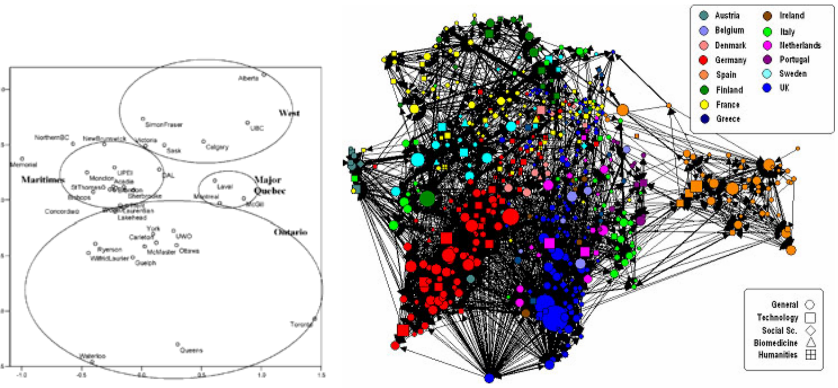

Figure 2. Co-link technique results: (left) Map of the Canadian (Vaughan 2006) and the (right) European (Ortega et al. 2008) academic web space.

Network graphs

Network modelling allows to visually represent the link structure of a space web according to the web unit used. It makes it possible to uncover structural properties of the nodes using social network indicators. Just as the Co-link Analysis, the network graphs may be generated through search engine and crawler data building a weighted matrix of directed links. However, the network can directly be plotted because it does not need any statistical processing but rather a network visualization program such as Pajek or NetDraw. There are several energizing algorithms (Kamada-Kawai or Fruchterman-Reingold), that optimize the graph visualization when it is complex and densely packed. These algorithms assume that the nodes are attracted or repelled according to their energy, which makes closer or farther the location of these nodes regarding to the number of links that they have. Kamada-Kawai algorithm is more suitable to small networks and only one component, while the

Fruchterman-Reingold one is appropriated to large networks and many components.

Both Network graphs and Co-link analyses allow adding properties to the nodes in order to observe relationships between the network configuration and other qualitative or quantitative variables. For example, the size of the nodes may represent the number of web pages, colour and shape any classification scheme such as country, type or discipline. These added variables permits to observe relationships between the centrality of a university and the number of web pages, or the colours make possible to identify national sub-networks in the World-class universities (Figure 3).

The network visualization has been mainly used at the level of web university domain, although there are others works at the web page level (Bjorneborn 2004; Cothey et al. 2005). Heimeriks and Van den Besselaar (2006) used it in order to detect four clusters in the EU-15 university web: German, British, Scandinavian and South European, while Ortega et al. (2008) observed that these clusters or national sub-networks are linked to the complete networks through prominent gateway universities.

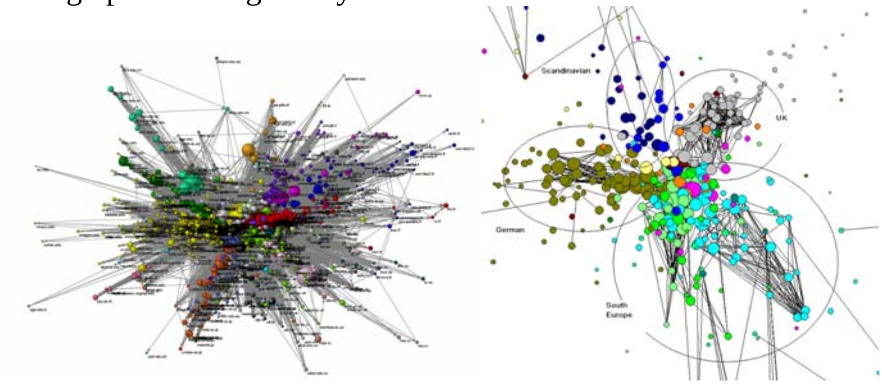

Figure 3. Network graphs of academic web spaces. (left) World-class universities on the Web (Ortega and Aguillo 2009), (right) EU-15 universities hyperlink network (Heimeriks and Van den Besselaar 2006).

Geographical maps

A third way to visualize web data is through a geographical metaphor. Geographical maps allow to present information at the macro level and assign web magnitudes to a certain region of the World. It makes possible observe geographical patterns in the web content and links distribution. To design a geographical map it is necessary two essential elements: a base map and data. The base map is an empty map where each region boundary is associated to an index in a database, while the data are grouped by regions and linked to that index as well. These maps are usually built using Geographical Information System (GIS) software which allows adding different layers, classification method and different map projections. Although multiple layers can be aggregated, it is recommended for simplicity to use only two, a hutch map which represents the number of web pages by region and a flow map which shows the links between those regions. There are several classification methods which distribute the data in classes (Standard deviation, Jenks' natural breaks, Percentiles, etc.), but the most usual and effective is the Jenks' natural breaks. This method determines the best arrangement of values into classes by iteratively comparing sums of the squared difference between observed values within each class and class means. This method improves the visualization and the interpretation of the results, because it creates more significant differences between classes.

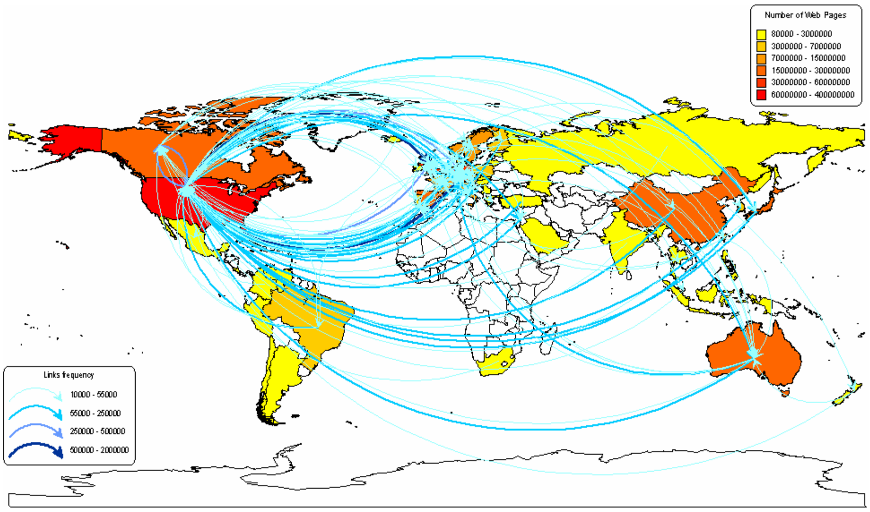

Figure 4. Geographical representation of web data. The 1000 most important universities on the Web grouped by countries and link flows (Ortega y Aguillo 2009).

Relevance of the university performance on the Web

Link analysis is a powerful tool for evaluating performance of institutions, especially those with a diverse group of stakeholders involved. Academic organizations are usually evaluated using peer review, consulting scholars or indirectly through bibliometric citation analysis. In both cases only colleagues are taken into account, ignoring the impact of other-than-research university missions. Even more important these scientometric techniques are excluding explicitly the economic, sociological, cultural or political impact of the academia and as already pointed out in many papers with strong biases against developing countries contribution.

External inlinks distribution to university websites provides a rich and diverse source of information about the visibility and impact of the university web presence. If this reflects the whole set of activities of the university, its global output, its performance according to its excellence and prestige then webometric indicators are the easiest and more powerful academic and research policy tool (Aguillo, 2009).

In order to take advantage of this situation but also for starting a virtuous circle as considering web indicators in academic evaluation will increase digital presence of universities, a Ranking of institutions was build using a composite webometric indicator. The Ranking Web has several technical advantages: Most of the universities have only one main web domain, so affiliation normalization is no longer a problem. The data is collected from the huge databases of the main search engines; with different geographic coverage but limited overlap among them. Both activity and impact can be computed from the number of webpages and documents and the number of external inlinks received in the university webdomain.

The Ranking Web of World Universities, also known as the Webometrics Ranking (www.webometrics.info) is published two times per year (January & July) since 2004 and analyzes the web presence of over 20,000 higher education institutions worldwide (Aguillo et al., 2008). Since 2006 the number of Open Access papers published is collected from the Google Scholar bibliographic citation database. During the last decades the ISI/Thomson databases were the only source for this information, being challenged only very recently by the SCOPUS/Elsevier database. Choosing a free alternative like Google Scholar is promoting many institutions publish their papers in web directories indexed by Scholar crawler increasing significantly the coverage and reducing the (still very relevant) biases and shortcomings of this Google database.

An ongoing global network analysis of the webometrics ranking results is showing both expected and unexpected results:

- There is an academic digital divide between North American universities that appear far better positioned than their European counterparts.
- US and Canadian universities are grouped together forming a unit in the Webspace, with French-speaking institutions not far from the core.
- European universities are split in several national or linguistic (Austria & Germany) groups not closely related to each other.
- In many countries a single university acts as a central gateway to the international academic network.

This global approach, considering the whole set of in and out-links, has many advantages as it allows to uncover relationships with other non academic stakeholders, but also a few shortcomings as probably some of the links coming from third parties are spurious. So, the future research on link networks will require classifying “a priori” the links according to their motivations (Wilkinson et al., 2003). Academic web networks could be cleaned allowing the identification of invisible colleges at a global level previously not achieved.

But the effort for identifying the different groups of motivations can be very high, so perhaps other alternatives could be also explored. According to link topology you can identify shallow links from deep links (Vasileiadou & van den Besselaar, 2006). The first ones refer to links between main homepages such as the central page of an academic institution. In this case the institutional interlinking could be drive by perception of prestige, the sense of community or common interests. An application for these links is to identify the pattern of out-links of a page and to build a set of webpages with a similar pattern. Using a quantitative approach you can use as a helping tool in search recovery like the related operator in Google or to visualize the neighborhood of an institution, as it is developed by Touchgraph (www.touchgraph.com) in the Figure 5.

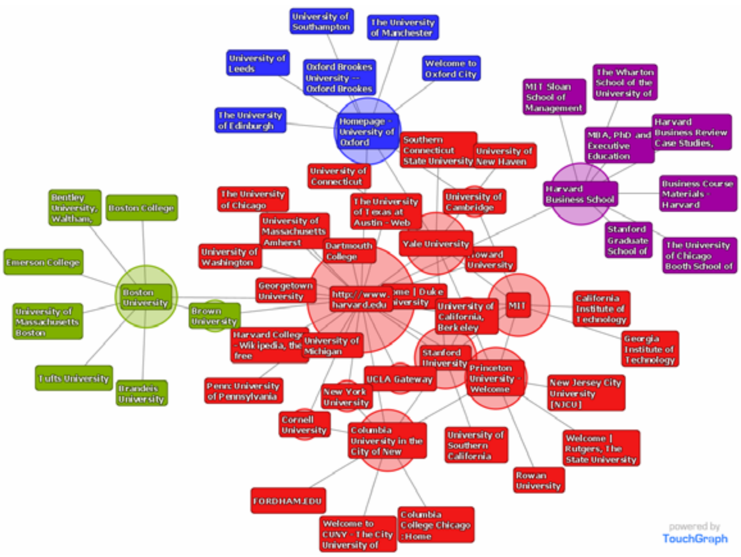
University of Southampton
The University of Manchester
University of Leeds
Oxford Brookes University -- Oxford Brookes
Welcome to Oxford City
The University of Edinburgh
Homepage - University of Oxford
Southern Connecticut State University
University of New Haven
MIT Sloan School of Management
The Wharton School of the University of
MBA, PhD and Executive Education
Harvard Business Review Case Studies,
University of Connecticut
Harvard Business School
Business Course Materials - Harvard
Bentley University, Waltham,
Boston College
The University of Chicago
The University of Texas at Austin - Web
University of Cambridge
Stanford Graduate School of
The University of Chicago Booth School of
University of Massachusetts Amherst
Dartmouth College
Yale University
Howard University
Emerson College
University of Washington
Georgetown University
http://www.harvard.edu
ome | Duke iversity
University of California, Berkeley
MIT
California Institute of Technology
Boston University
University of Massachusetts Boston
Brown University
Harvard College - Wikipedia, the free
University of Michigan
Stanford University
Georgia Institute of Technology
Princeton University - Welcome
UCLA Gateway
New Jersey City University [NJCU]
Tufts University
Brandeis University
Penn: University of Pennsylvania
New York University
Cornell University
Columbia University in the City of New
University of Southern California
Welcome | Rutgers, The State University
Rowan University
FORDHAM.EDU
Welcome to CUNY - The City University of
Columbia College Chicago :Home
powered by TouchGraph

Figure 5. Neighborhood of the Harvard University according to its out-link pattern, showing affinities with other Ivy League universities and international relationships with UK universities

Deep linking refers to links among the contents published in the web directories and they are important in academia as the websites gets richer and more diverse allowing building complex networks and answering new questions. A short list of some of these questions is proposed for future research:

- Can be estimated the relative contribution of each of the university missions to the global web performance of the university? What are the reason explaining possible discrepancies between obtained and perceived results?
- What it is importance of the disciplines in the self-organizations of web networks, specially targeting the problems related to humanities and social sciences?
- According to central measures, what is most relevant for the web domain, the formal or the informal scholarly communication processes and outputs?
- What is the impact of Web 2.0? How is publish, used, linked the media contents?
- What it the relative contribution to the Webspace of the non-academic activities? For example, is it the Ivy League today a group of elite universities or mainly a sports league?
- Are there technical or information guidelines applied correctly and what is the impact of the web bad practices?

References

Aguillo IF, Granadino B, Ortega JL, Prieto JA (2006) Scientific research activity and communication measured with cybermetrics indicators. Journal of the American Society for Information Science and Technology 57(10): 1296 – 1302

Aguillo, I.F.; Ortega, J. L. & Fernández, M. (2008). Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments. Higher Education in Europe, 33(2/3): 234-244.

Aguillo, I. (2009). Measuring the institution's footprint in the web. Library Hi Tech, 27 (4): 540-556.

Antelman K (2004). Do Open-Access Articles Have a Greater Research Impact? College and Research Libraries 65(5): 372-382.

Almind TC, Ingwersen P (1997) Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. Journal of Documentation 53 (4): 404–426.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509-512

Bar-Ilan J (2005) Expectations versus reality – Search engine features needed for Web research at mid 2005. Cybermetrics 9(1) http://cybermetrics.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

Bar-Ilan J (2008) Which h-index?—A comparison of WoS, Scopus and Google Scholar. Scientometrics 74(2): 257-271

Bar-Ilan J, Peritz BC (2009) The lifespan of 'informetrics' on the Web: An eight year study (1998-2006). Scientometrics 79(1): 7-25.

Bianconi G, Barabasi AL (2001) Competition and multiscaling in evolving networks. Europhysics Letters 54, 43M42.

Björneborn L (2004) Small-world link structures across an academic web space: a library and information science approach. Dissertation, Royal School of Library and Information Science

Björneborn L, Ingwersen P (2004) Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology 55 (14): 1216–1227

boyd dm, Ellison NB (2007) Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13 (1): article 11. http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30 (1-7): 107-117.

Broder A, Kumar R, Maghoul F et al (2000) Graph structure in the Web. Computer Networks 33(1-6): 309-320.

Chakrabarti S (2002) Mining the Web: Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers, San Francisco

Chen H, Shen H, Xiong J et al (2006) Social Network Structure behind the Mailing Lists: ICT-IIIS In: TREC 2006 Expert Finding Track, Gaithersburg, Maryland

Cho J, Garcia-Molina H (2000) The Evolution of the Web and Implications for an Incremental Crawler. In: Proceedings of the 26th International Conference on Very Large Data Bases, San Francisco

Cothey V, Aguillo IF, Arroyo N (2006) Operationalising "Websites": lexically, semantically or topologically? Cybermetrics**,** 10(1): Paper 4. http://www.cindoc.csic.es/cybermetrics/articles/v10i1p4.html

Fetterly D, Manasse M, Najork M et al (2003) A large-scale study of the evolution of web pages. In: Proceedings of the 12th International World Wide Web Conference, ACM Press, Budapest,

Gomory S, Hoch R, Lee J et al (1999) Analysis and Visualization of Metrics for Online Merchandizing. In: Hochheiser H, Shneiderman B (ed) Understanding Patterns of User Visits to Web Sites: Interactive Starfield Visualization of WWW Log Data. Springer, San Diego

Harter S, Kim H (1996) Electronic journals and scholarly communication: a citation and reference study. Information Research 2(1): paper 9a. http://informationr.net/ir/2-1/paper9a.html

Harter SP, Ford CE (2000) Web-based analyses of E-journal impact: Approaches, problems, and issues. Journal of the American Society for Information Science 51(13): 1159-1176

Heimeriks G, Van Den Besselaar P (2006) Analyzing hyperlinks networks: The meaning of hyperlink based indicators of knowledge production. Cybermetrics 10(1,1). http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.html

Jansen BJ, Spink A, Pederson J (2005) A Temporal Comparison of AltaVista Web Searching. Journal of the American Society for Information Science and Technology 56(6): 559–570.

Kleinberg J (1999) Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5): 604-632.

Koehler W (2004) A Longitudinal Study of Web Pages Continued: a Consideration of Document Persistence. Information Research 9(2): paper 174 http://informationr.net/ir/9-2/paper174.html

Kretschmer H, Aguillo IF (2005) New indicators for gender studies in web networks. Information Processing & Management 41(6): 1481 - 1494

Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, Philadelphia

Larson R (1996), Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of Cyberspace, In: Proceedings of ASIS96, 71-78. http://sherlock.berkeley.edu/asis96/asis96.html

Lawrence S, Giles CL (1998) Searching the World Wide Web. Science 280(5360): 98-100.

Lawrence S, Pennock DM, Flake GW et al (2001) Persistence of Web References in Scientific Research. Computer 34(2): 26-31.

Lenhart A, Madden M (2007) Teens, privacy, & online social networks. Washington, DC: Pew Internet and American Life Project Report. http://www.pewinternet.org/~/media//Files/Reports/2007/PIP_Teens_Privacy_SNS_Report_Final.pdf.pdf Accessed 8 July 2010

Ortega JL, Aguillo IF (2008) Visualization of the Nordic academic web: Link analysis using Social Network tools. Information Processing & Management 44(4): 1624-1633

Ortega JL, Aguillo IF (2008) Linking patterns in the European Union’s Countries: geographical maps of the European academic web space. Journal of Information Science 34(5): 705-714

Ortega JL, Aguillo IF (2009) Mapping World-class universities on the Web. Information Processing & Management 45(2): 272-279

Ortega JL, Aguillo IF (2010) Differences between web sessions according to the origin of their visits. Journal of Informetrics 4(3): 331-337

Ortega JL, Aguillo IF, Cothey V, Scharnhorst A (2008) Maps of the academic web in the European Higher Education Area — an exploration of visual web indicators. Scientometrics 74(2): 295-308.

Ortega JL, Aguillo IF, Prieto JA (2006) Longitudinal Study of Contents and Elements in the Scientific Web environment. Journal of Information Science 32(4): 344-351.

Ortega JL, Cothey V, Aguillo IF (2009) How old is the Web? Characterizing the age and the currency of the European scientific Web. Scientometrics 81(1): 295–309.

Paine N, Thelwall M (2008) Do academic link types change over time? Journal of Documentation 64(5): 707-720

Pennock DM, Flake GW, Lawrence S et al (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.

Pitkow J (1997) In search of reliable usage data on the WWW. In Sixth International World Wide Web Conference Santa Clara, CA, (pp. 451–463).

Rodriguez Gairín JM (1997) Valorando el impacto de la información en Internet AltaVista, el "Citation Index" de la Red. Revista Española de Documentación Científica 20(2): 175-181.

Rousseau R (1999) Daily time series of common single word searches in AltaVista and NorthernLight. Cybermetrics 2/3: Paper 2. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Silverstein C, Henzinger M, Marais H, Moricz M (1998) Analysis of a Very Large AltaVista Query Log. SRC Technical note, 14. http://citeseer.ist.psu.edu/70663.html

Smith A, Thelwall M (2002) Web Impact Factors for Australasian universities, Scientometrics 54(1-2): 363-380.

Sullivan D (2003) Google Dance Syndrome Strikes Again. SearchEngineWatch.com, http://searchenginewatch.com/showPage.html?page=3114531 Accessed 8 July 2010

Thelwall M (2001) Extracting macroscopic information from web links. Journal of the American Society for Information Science and Technology 52(13): 1157-1168.

Thelwall M (2002) Evidence for the existence of geographic trends in university web site interlinking. Journal of Documentation 58(5): 563-574

Thelwall M (2008) Social networks, gender and friending: an analysis of MySpace member profiles. Journal of the American Society for Information Science and Technology 59(8): 1321–1330

Thelwall M, Klitkou A, Verbeek A et al (2010) Policy-relevant webometrics for individual scientific fields. Journal of the American Society for Information Science and Technology 61(7): 1464-1475.

Teevan J, Adar E, Jones R, Potts M (2006) History repeats itself: repeat queries in Yahoo's logs. In: Proceedings of the 29th Annual international ACM SIGIR, Seattle

Valverde S, Ferrer-Cancho R, Sole RV (2002) Scale Free Networks from Optimal Design. Europhys. Lett. 60: 512-517.

Vasileiadou E, van den Besselaar P (2006) Linking shallow, linking deep. How scientific intermediaries use the Web for their network of collaborators. Cybermetrics 10(1): paper 4. http://www.cindoc.csic.es/cybermetrics/articles/v10i1p4.pdf

Vaughan L (2006), Visualizing linguistic and cultural differences using Web co-link data. Journal of the American Society for Information Science & Technology 57(9): 1178-1193.

Vaughan L, Thelwall M (2003) Scholarly use of the web: What are the key inducers of links to journal web sites? Journal of the American Society for Information Science and Technology 54(1): 29-38.

Wang W, Zaiane OR (2002) Clustering Web Sessions by Sequence Alignment. Proceedings of 13th Int. Workshop on Database and Expert Systems Applications, Aix-en-Provence.

Wilkinson D, Harries G, Thelwall M et al (2003) Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. Journal of Information Science 29(1): 49-56