Exploring web navigational paths through social network analysis

José Luis Ortega[1]
VICYT-CSIC, Madrid, Spain

Isidro F. Aguillo
Cybermetrics Lab, CCHS-CSIC, Madrid, Spain

Abstract

Purpose: The aim of this paper is to present the network visualization and SNA indicators as a way to explore and to analyse the web site design through the users' navigational paths.

Design/methodology/approach: Web log files from the webometrics.info site were analysed. Web usage mining techniques were used to process these data. Several network graphs (navigational and query graph) were built to inspect the latent structure from the user view. SNA techniques and indicators were used to extract the main results.

Findings: Results show networks with small-world and scale-free properties; the double structure of the two language versions; the central position of the search engines as hubs that distribute the access to every page; how the pages are grouped by their thematic and structural relationships; and which pages are less requested through search engines queries or are less visible.

Research limitations: Results are based on the web performance of only a web site and its generalization has to be considered with caution.

Originality/value: The application of network visualization and SNA indicators allows to explore the web site navigation by users and to test how it is used. The graph representation makes possible an intuitive way to observe the structure of a web site, the behaviour of the users and the information consumption.

Introduction

The study of the usage that a web surfer make of a web site, the pages that they visit and how they behave is one of the most important questions that the online marketing sector needs to answer in order to fulfil the customer's demands (Lee et al., 2000; da Cruz et al., 2004). From the *market basket analysis* in which association rules are applied to relate products among them and between client profiles, the study of the user needs has growth more and more due to the existence of huge amount of information on the user transactions, and to the incessant need to know the behavioural habits of the customers. In this sense, web transaction logs, where the footprint of the user's web navigation is engraved, are the quantitative registers that nowadays most information gives about the behaviour of the website users and their informational needs (Catledge and Pitkow, 1995). As the shopping ticket, web log files constitute a record of the user's activity where it is able to observe information consumption, user groups, tasks achievement and design failures. Beside to the user tests (Nielsen, 1993), the web usage mining is the only way to understand the needs of the users in a web environment and to check the running of a web site (Cooley et al., 1997; 1999).

The visual representation of quantitative data makes possible to obtain a fast understanding of the displayed reality. It may help to the patterns detection and to the

[1] VICYT-CSIC, Serrano, 113, 28006, Madrid, Spain, jortega@orgc.csic.es

systematization of complex relationship among various data types (Chen, 2004). Regarding to the web log analysis, network graphs allow to present frequent navigational paths, web pages clusters and topological structures of a web site which makes possible the decision taking on the management of web sites, search engine optimization (SEO) strategies and future development lines. The web site structure from the view of the users, through their clicks, shows how they access to a site and explore the contents. However, these visualizations would be just a graphical view without the support of the social network analysis (SNA) techniques (Nooy, Mrvar and Batagelj, 2005). These measurements define as the global structure of the network as the value of each webpage into it, which helps to understand how a site is used.

Related research

The huge size of web log files and the need to process and to clean this data before a satisfactory analysis, demands the employment of advanced data processing techniques. Web usage mining (Cooley et al., 1997; 1999), a type of data mining focused in the web log files analysis, provides tools to analyse these files, making possible to identify and locate the users, tracking each web session length and duration (Pitkow, 1997; He et al., 2002), forecasting the user navigation path (Nanopoulos, et al., 2001; Yang, et al., 2002) and automatically improving the organization and presentation of websites (Perkowitz and Etzioni, 1998; 1999). This processing technique has been widely used to group web pages according to the navigational preferences of the users, disclosing semantic relationships between them (Zhu et al., 2004) and introducing new clustering methodologies (Cadez et al., 2003).

Visualization of the web site structure as a way to contents management and design improvements was already suggested in 1994. Pitkow and Bharat (1994) proposed WebViz as a 3D graphical tool in order to represent the hierarchical structure of a web site pointing visited web pages. From this part, several works have presented different visualization proposals. Munzner and Burchard (1995) introduced a hyperbolic space in the website structure representation, while Frecon and Smith (1998) developed Webpath, an application that used the three dimensions to display the browsing history. Hang and Landay (2001), through the WebQuilt software, presented a graphical way to rebuild the route of the most frequent paths for a given task. Erbacher (2001; 2002) presented glyph-based network graphs to monitoring the web traffic. More actually, visual software tools and data mining algorithms are integrated in web management packages. Minarik and Dymacek (2008) suggested NetFlow Visualizer as an integrated pack to visualize the traffic flows of an organization at different levels. Other works are more focused in the utility of the visualization as a way to improve the website usage, than in the design of visual metaphors. Thus, Chi (2002) utilized the traffic visualization as a way to detect web design anomalies, and Ting et al. (2004) developed visual graphs to detect and classify browsing patterns. Meanwhile, other authors explored the design aspect of this type of graphs (Fry, 2004).

SNA has been used to study relational properties of the web pages. Different authors have analysed the relationship between the web architecture and the clicks network, finding that navigational process does not follow a random walk as the PageRank assumes (Meiss et al., 2008), while others claim that the topology of a web site is very instrumental in guiding the users through the site (Borges and Levene, 2006). Several structural models such as the Markov chains (Eirinaki and Vazirgiannis, 2007), learning automata (Forsati and Meybodi, 2010) or other probabilistic models (Chakrabarti et al., 2009) have been applied to study the users browsing on a website in order to improve their functionalities or to plan recommendations services. However, although some

works have applied social network measures (Gloor and Zhoa, 2004), there are no studies that systematically employ social network indicators beside to network visualizations.

Objectives

This work intends to show the utility of the social network analysis (SNA) and its visualization as a diagnostic tool for the study of web usage data. Thanks to the visualization of paths through web sites we can rebuild the web structure according to the users' behaviour, to detect design flaws and to suggest new ways to develop the web service. We propose adding this type of analysis to the evaluation of website performance and the SEO strategies. The following research questions are proposed:

- Is it possible a systematic use of the SNA indicators on web log files as a way to analyse and to understand the main topological characteristics of a navigational network?
- Is the visualization of the browsing patterns a suitable inspection method to describe the use of a web site and to complement the structural analysis?
- Could this visual graphs and network measures be important tools that support the decision making in web management and SOE strategies?

Research Design

Web log transactions processing

The Ranking Web of World Universities (webometrics.info) is a portal that ranks 12,000 (3,000 in 2006) universities according to two main criteria: size (number of pages and rich files) and visibility (number of incoming links). It is the most complete and updated ranking of universities web domains. This website is very popular with more than 4 million visitors per year and a Page Rank of 8 (Figure 1). We think that the high visibility of this site provides a good sample to study the access pattern of a website. Web log transactions from 2006 July were selected as a sample to carry out the web usage analysis. This web log file is in NCSA Combined Log Format which included information on the URL accessed and the referent page. A data cleaning process was done to remove no relevant accesses. The deleted visits were:

- To graphic files (gif, jpg and png)
- To style sheets (css)
- Which do not request a petition (get)
- From the own website editor IPs (161.111.200.*)
- Made by crawlers or bots (Googlebot, Msnbot, Slurp, Gigabot, etc.)

Figure 1. Snapshot of the main page of The Ranking Web of World Universities (webometrics.info) in 2006 June

To generate the navigational network, only two fields of this file were used: the request and the referrer fields. The referrer, which contains the web page from which the access was made, was used as starting point, while the requested URL was recorded as end point. This allows to build a direct network that represents the browsing flows through the website. Thus, this weighted network of 276,587 visits contains 10,766 distinct accesses to 1,149 webometrics.info's web pages from 725 external web sites. In order to reduce the complexity of the network visualization a cut-off level of 2 accesses was used which reduced the network to 6,479 arcs.

In the case of the queries network, we have used the query chain contained in the visit from search engines as registered in the referrer field of the web log transactions. We have selected the complete sentences instead of individual terms because these do not show the specific way in which a user locates a webpage and because the ranking of web pages in search engines are sensitive to the use of different variants in a query (Bar-Ilan, 2005). 55,766 queries submitted to the search engines started a visit to

webometrics.info site. The two most popular search engines used to visit the sample site were Google (including all national versions) with a 46.8% and Yahoo! Search with the 4.7% of the queries. We also have used a cut-off level to reduce the complexity of the network and we have selected the queries with a frequency equal or higher than 5. Then the resulting weighted network contained 547 distinct queries that access to 790 web pages. To visualize the graph, a network analysis software package was used, Gephi 0.7 beta version with GNU licence and developed by Gephi Consortium (consortium.gephi.org). Finally, the layout algorithm used to energize the network was a force-directed algorithm (Force Atlas) which is implemented in this software (Bastian, Heymann and Jacomy, 2009). This layout algorithm was used because is specialized in large scale-free networks, converging more efficiently and increasing the speed in the network configuration.

Structural indicators

Several network indicators were calculated to measure the principal network characteristics and to know the specific relationships of each node in the network. Gephi 0.7 was used to calculate these metrics. These indicators are:

- Degree centrality (k): It measures the number of lines incident to a node (Freeman, 1979). A variation is the weighted centrality degree which calculates the weight of each line, indicating the strength of each relationship. In this study the centrality degree allows to describe the path that a user follows exploring the website, indicating the multiple access points of a page and its relationship with other pages. Given that this is a directed network, the centrality degree can be expressed in two opposite ways, InDegree and OutDegree. InDegree shows the number of incoming visits to a web page from distinct pages while the OutDegree indicates the number of exit visits from a web site to different pages. Indegree is a good index to detect the main gateways to a site, pointing the pages with highest interest and therefore with the most relevant content. Meanwhile the Outdegree identifies the links to the main external sites and the key navigational pages that guide the site browsing. In Gephi, it is calculated in menu Window > Statistics > Average Degree.
- Freeman's Betweenness centrality (CB): It is defined as the capacity of one node to help to connect those nodes that are not directly connected between them (Freeman, 1980). This measurement enables us to detect important hubs that connect different web contents being these hubs critical pages to the site browsing. This is calculated in Statistics > Network Diameter.
- Global clustering coefficient (C): It is a measure that shows the density or cohesion of a network. It shows the proportion of nodes that tend to group together. Mathematically, it is the proportion of closed triads by open triads; being a triad a group of three linked nodes. This measure is important to detect 'small world' phenomena. It can be obtained in Statistics > Avg. Clustering Coefficients > Directed. To calculate the Clustering Coefficient to a random network is just necessary to divide the average number of arcs per node by the total number of nodes in the network.
- Average path length (Distance): It is the number of steps in the shortest path that connects two nodes. The average number among of all the shortest paths in the network is the average distance. A low mean average distance is a good indicator of network density. This informs about the accessibility of the contents in the network and the click distance between all of the pages analysed. It is obtained in Statistics > Avg. Paths Length.

Results

The resulting navigational network shows small-world properties as its clustering coefficient (C=.194) is slightly higher than the expected for a random network (C=.003) although its average path length (l=4.06) is slightly high (Watts & Strogatz, 1998). Both the centrality Indegree (γ=-4.32) and Outdegree (γ=-5.19) frequencies distribution follow a power law, a trend which allows us to state that this network owns scale-free properties as well (Barabasi, Albert & Jeong, 2000). The slope coefficients (γ) were calculated from Gephi 0.7. They can be extracted from Statistics > Degree Power Law > Directed.

Navigational graph

Figure 2 shows the network of clicks that the webometrics.info's users followed in 2006 July. Each web page is coloured by their section on the web site architecture: for example, the light green nodes represent pages about the "Universities by country" while the blue ones belong to the "Top 100 universities by continent". The visits that come from external sites are coloured in salmon. The size of each webpage represents the number of times that it is visited in the case of internal pages. The size of external sites is related with the number of times that links to a webometrics.info's webpage.

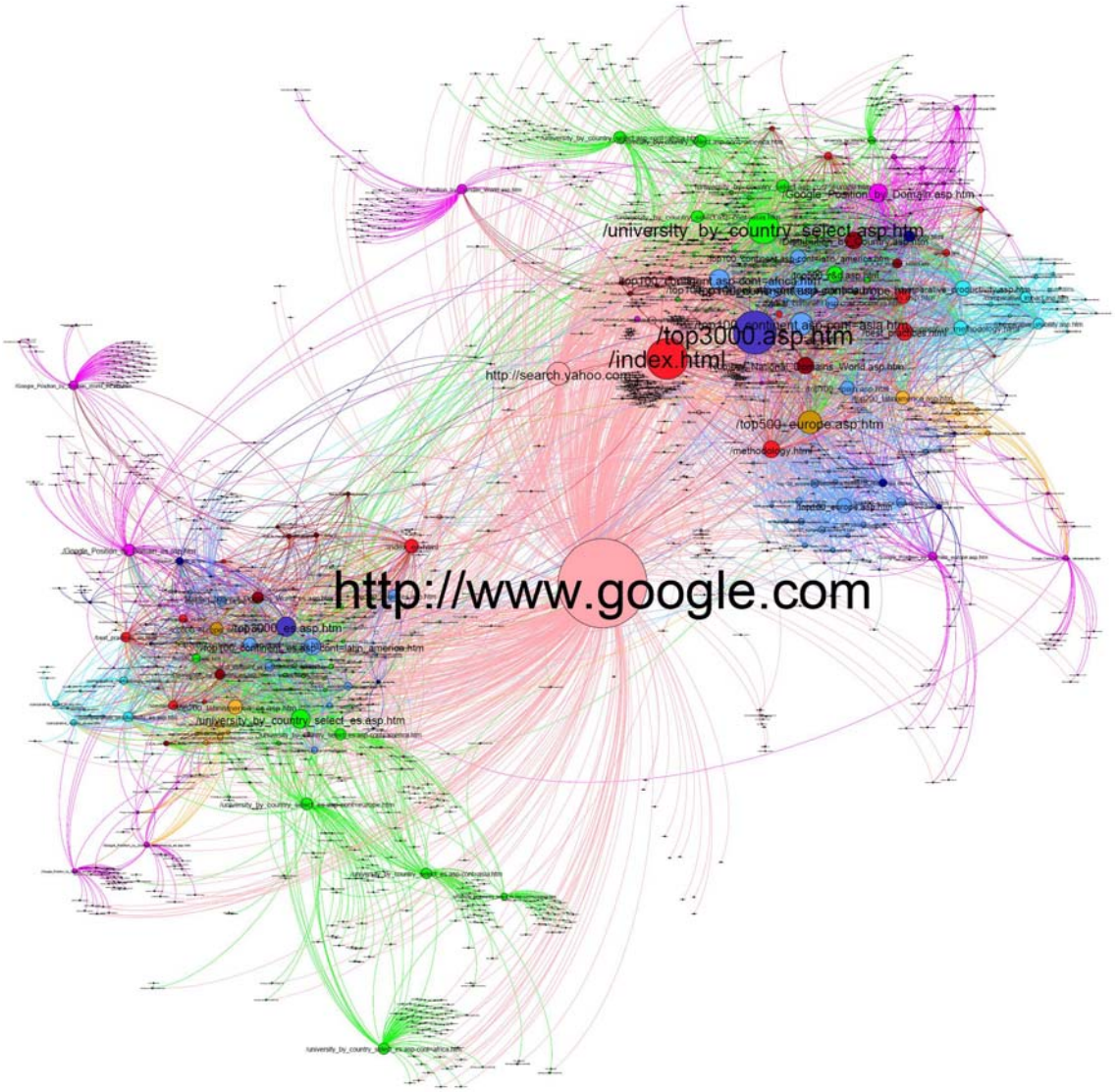

Figure 2. Network of clicks between internal and external web pages of webometrics.info

One of the first facts that we can observe in the Figure 2 is that the web site is split in two clearly segregated sections: the English language part (upon right) and the Spanish language one (bottom left). It shows that the users that visit the English version scarcely visit the Spanish one, partly probably because there are very few links between both versions. In fact, both parts of the website are linked through external websites; mainly search engines services such as Google which links to 651 (56%) webometrics.info's webpages. The size of the nodes in the English part is larger than the Spanish one due that the English version is the most visited pages with the 82% of the web traffic.

The graph also shows which sections are more visited from external links or search engine queries. Thus, the closest pages to the centre (to the search engines sites) are those mainly accessed from search engines such as the main pages (red), the "Top 100 universities by continent" (blue) and the "Universities by country" ones (light green). Meanwhile, the pages less visited from search engines are located far from the central positions such as the "Comparative Analysis" pages (light blue) and the "Google position by domain" ones (pink). This pattern is similar in both language versions. This result informs us that those pages may be located in deeper places with a lower visibility

for search engines or that they are not interesting enough for the users who do not search for them.

Page	*In Degree*	*Out Degree*	*Degree*	*Weighted Degree*
/top3000.asp.htm	277	32	309	118,820
http://www.google.com	0	651	651	25,982
/top3000_es.asp.htm	109	26	135	17,165
/index.html	263	32	295	13,445
/top100_continent.asp-cont=europe.htm	120	28	148	12,338

Table 1. Rank of the pages with the most centrality degree.

Page	*Betweenness*
/university_by_country_select.asp.htm	306,966
/university_by_country_select_es.asp.htm	270,037
/index.html	182,954
/top3000.asp.htm	118,747
/top500_europe.asp.htm	110,372

Table 2. Pages with highest Betweenness centrality

Table 1 presents the five web pages with the highest centrality degree. The most visited pages are "/top3000.asp.htm" and "/top3000_es.asp.htm", both English and Spanish version of the main ranking "Top 3000 Universities", followed by the "/index.html" page and the "/top100_continent.asp-cont=europe.htm" ("Top Europe"). However, the pages with the highest InDegree (visits from different pages) are "/top3000.asp.htm" (k_{in}=277) and "/index.html" (k_{in}=263) which can be considered the main gateways to the website and the pages with the most popular contents. According to the OutDegree, the sites that link to most of the webometrics.info's pages are the two main search engines, Google (k_{out}=651) and Yahoo! Search (k_{out}=162), while the internal pages "/university_by_country_select.asp-cont=europe.htm" ("Universities by country"> "Europe") with 75 and "/university_by_country_select.asp-cont=asia.htm" ("Universities by country" > "Asia") with 73 are the pages that most traffic distribute to other pages. With regard to the Betweenness centrality (Table 2), the pages with a highest value are the navigational pages "/university_by_country_select.asp.htm" (*CB*=306,966) and "/university_by_country_es.asp" (*CB*=270,037) because these pages share out the traffic to deep content pages about rankings of universities by country.

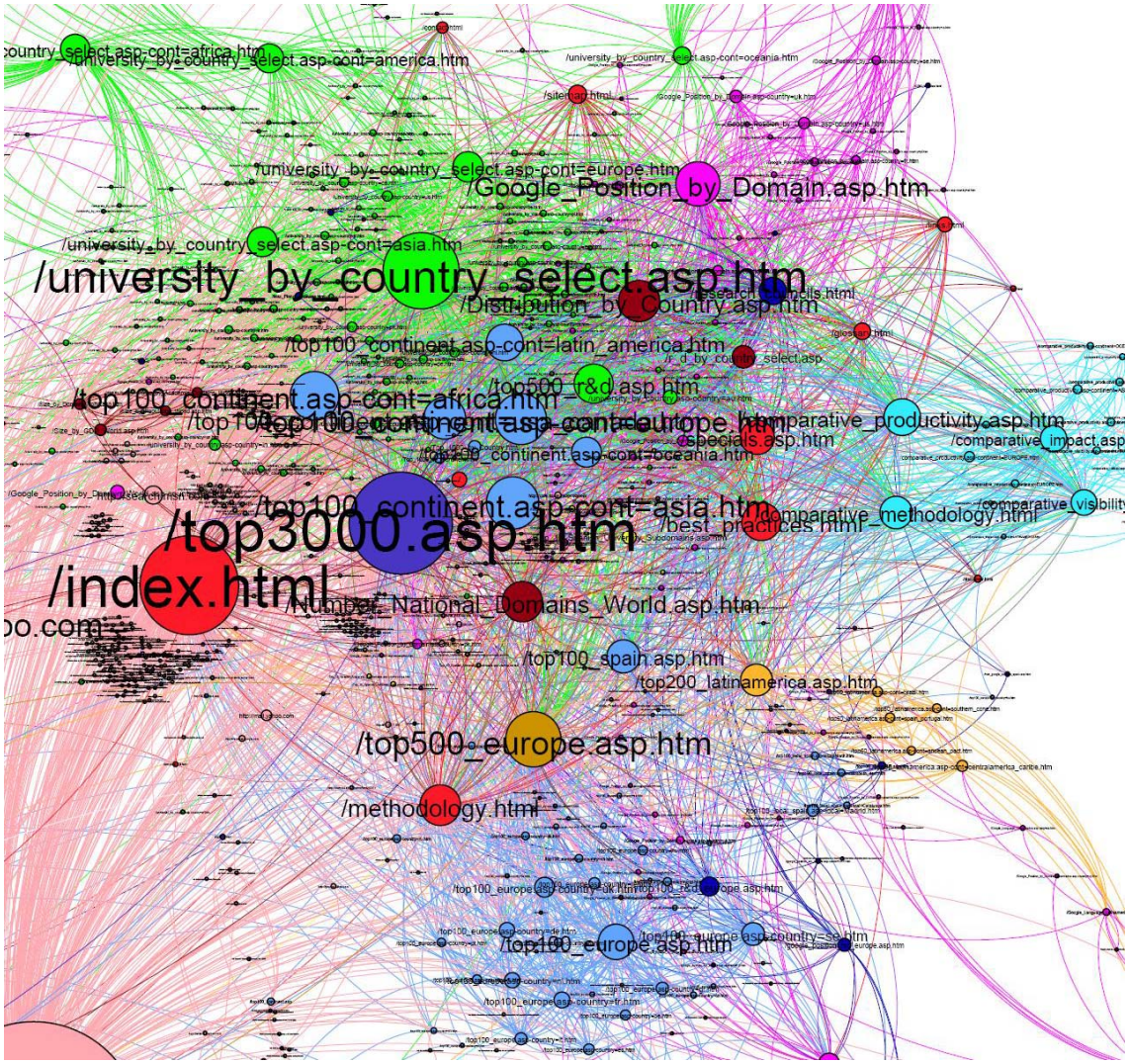

Figure 3. Detail view of the English version website of webometrics.info

The Figure 2 also lets us to appreciate how the webpages are clustered with the same pages of their section. Thus, we can observe in detail the English part (Figure 3) where we detect in blue the "Top 100 universities by continent" pages; in light blue the "Comparative Analysis" section and in light green the "Universities by country" pages. However, we can found webpages that are located far away from their sections and they are integrated in other sets. For example, the bottom cluster "Top Europe" (in blue) is also set up by "Google position by Domain" (dark blue) and "European Rank" (sandy), setting up a group of web pages about Europe. This group consists of pages about Europe even although they belong to different sections which it is advisable to group these pages in the same geographical category. This misclassification is also observed with other Top100 pages such as "Spain rank" pages or "Top 100 universities by continent" ones which are grouped in different places. This lets us to suggest that this category would be reclassified in two different sections: Top 100 continent and Top 100 Europe. Hence, the visualization of this clicks network makes possible to observe how the navigation flows describe the topology of our website and to detect anomalies and improvements in its design through a visual inspection method.

Query graph

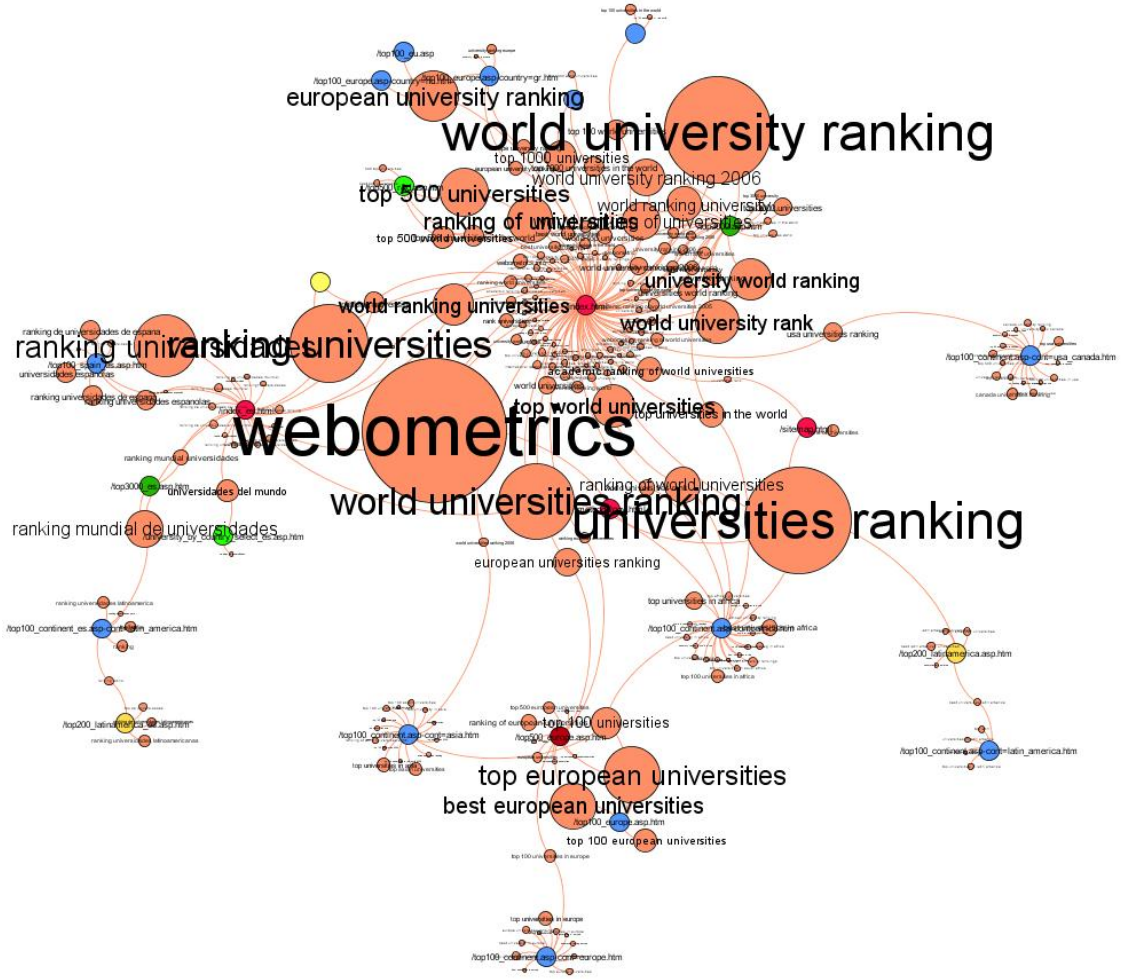

Figure 4. network of webpages visited through the most frequent queries

Query	*Frequency*	*%*
Webometrics	260	1.16
world university ranking	191	0.85
universities ranking	190	0.85
ranking universities	139	0.62
world universities ranking	135	0.60
Total	22,380	100

Table 3. The 5 most frequency queries

A different point of view of the usage that the visitors make of a website is related with the queries that they send to a search service to locate web sites which may be of interest to them because this way of surfing is the most usual (Lavene, 2005). The modelling of this network of queries and visited pages allows to know what queries are the most usual and which find certain web pages. Figure 4 presents only the giant component of this network of queries and pages composed of 267 (11%) nodes and 307 (58%) edges. It shows that the almost half of the queries (47%) link to the index pages, which suggests that the main pages are the principal access points to the web site. If we analyse the most frequent queries (Table 3) we can observe that they are referring to the title of the site but not to the content which causes this high access to the main pages.

Excluding the main pages, the second group of most requested pages is the "Top 100 universities by continent" pages, mainly, the European ones (bottom) which reflects the European interest for this site. Figure 4 also shows, in its left side, the Spanish queries that are commonly used to access to the Spanish version site, confirming that this part of the web site is less visited and it is less attractive than the rest of the site. It is interesting to notice that the five most frequent queries do not unequivocally looking for this site because webometrics is the name of a research discipline and the remaining four asking for information on universities rankings in general (including other universities rankings such as Shanghai's ARWU or QS World University Rankings). This could be due to the good position of this site on the search engines which are responsible of a large proportion of the web traffic about these issues.

Discussion

The employment of social network measurements makes possible not only to observe the main topological characteristic of the navigational network, but also to identify important web pages in the browsing process. From the topological view, the navigational network presents small-world properties because its average clustering coefficient is larger than the same for a random network (C=.194). This means that there are transversal links that directly relate remote pages. This characteristic may emerge from the search engines sites and the main pages of the site which link the great majority of web pages, increasing the connectivity and web traffic. The skewed distribution of the degree centrality indices confirms the presence of scale-free network properties as well. According to this, there are few pages that receive most of the traffic such as "/top3000.asp.htm" and "/index.html", which would be able to consider as navigational pages because distribute the web traffic to the rest of pages. Meanwhile, the remaining web pages, with regional/national rankings or information about each university, scarcely count on a couple of visits due to they are the pages with the largest contents and are located in the depth of the site.

According to the role or importance of each page in the network, pages with a high indegree centrality may be consider as the main gateways to the website and the pages with the most popular contents, while the most outdegree pages are mainly navigational pages that share out the traffic to content pages. Betweenness centrality shows the internal gateway pages that mediate between contents pages and navigational pages. These indicators allow to measure and to point the structural property of each node and precise if they conform to the assigned tasks in the web architecture besides to identify key pages in the web design.

The graphical representation comes to complement this structural analysis, offering a visual view of the navigational map with the addition of qualitative variables that enrich the picture. Through a visual inspection, important aspects of the web site structure and the browsing habits have been observed such as the double structure of the two language versions; the central position of the search engines as hubs that distribute the access to every page; how the pages are grouped by their thematic and structural relationships; and which pages are less requested through search engines queries or are less visible. Unlike other studies, in which the visualization are focused in the users behaviour (Hong and Landay, 2001; Herder and Weinreich, 2005) or in the technical aspect of large log files visualization (Chen et al., 2004; Youssefi et al., 2004), the navigational map of this study is used as a way to explore the main users surfing patterns, the topological structure of the site and the design faults. In this sense, the works of Chi et al. (2000) and Chi (2002) are in line with our study because they interpret the visualization as a tool for decision making. Through longitudinal view of

web sessions they extracted conclusions on the success or failure of certain web designs in Xerox web. Ting et al. (2004) implemented a recommendation system by means of the visualization of web patterns.
On the other hand, the query graph has shown that the principal gateways to access to the site are the indexes pages perhaps due to the most frequent queries ask for the title of site. However, there are few queries that search for specific sections. In this sense, we could highlight the queries about the European ranking or the Spanish section.

Limitations
Nevertheless, one of the most important limitations of the analysis of the web log files is its difficult to generalize its results. The use of webometric.info files is just an example which leads to build the graph and show their characteristics and particularities as a way to emphasize the potential of this type of analysis. Due to this, our results can not be extensive to other websites but indeed the use of this technique with other web site transaction log files, which would shed light on new functionalities of this technique and new uses of the web sites. Other important limitation is the availability of this type of data which are not public access and they are withheld by web system administrators (Bar-Ilan, 2007). In our case we only had at our disposal the June 2006 log. We think that although it is a little old because the web site have changed since that moment, is enough to show a example to propose the social network measurements as a powerful web usage mining technique.

Implications
The application of these structural measures and the visual representation of web accesses can be a suitable tool to the design of web management policies and to develop SEO strategies. It can be used as a new analysis tool to web managers and designers which can detect architecture gaps and malfunctions of the web applications. From a longitudinal view, consecutive graphs can be observed to appreciate the evolution of the users' behaviour and to confirm the utility of changes in the web structure. The query graph introduces the possibility to elect the most precise terms that reach the site in search engine queries and to adopt advertising policies.

Conclusions and Further Research
The implementation of the SNA indicators have made possible to describe the topological structure of the navigational network and to observe the main structural characteristics, which has allowed to define the role of each type of page in the network and to explain how the browsing flows are distributed across the web site. We can conclude that the systematic use of these measurements not only is a suitable proposal but also would be advisable in order to describe the latent structure of a web site and to suggest new improvements.
The results suggest that the use of network visualizations is an appropriate and recommended inspection method to explore the behaviour of the users of our web site. The navigational graph has shown a picture of the website structure through the route of the users across the web pages. Furthermore, this graphic representation has made possible to suggest improvements and to detect anomalies such the reclassification of the "Top 100 universities by continent" pages or to state the secondary role of the Spanish language version. This allows to conclude that both the navigational and the query graphs are suitable methods to describe the use of a web site and to complement the structural analysis.

We conclude that the use of network graphs and SNA techniques could be an important and valuable tool to help in the decision making and advising process on website construction, design and management. These techniques are highly recommended due to their simplicity to data processing and the understandability of their results which handily permit to interpret how is running a web site. Finally, the query processing and graphing would be a good tool to the study and analysis of the user needs and how they come into a web site as well as to help to the online marketing and SEO strategies.

References

Barabasi, A. L., Albert, R. and Jeong, H. (2000), "Scale-Free Characteristics of Random Networks: the Topology of the World-Wide Web", *Physica A*, Vol. 281No. 1-4, pp. 69-77.

Bar-Ilan, J. (2005), "Comparing rankings of search results on the web", *Information Processing and Management*, Vol. 41 No. 6, pp. 973-986.

Bar-Ilan J. (2007), "Position Paper: Access to Query Logs – An Academic Researcher's Point of View " in: *Log analysis workshop, WWW2007*. http://www2007.org/workshops/paper_39.pdf (accessed 04 November 2011)

Bastian, M., Heymann, S. and Jacomy, M. (2009), "Gephi: An Open Source Software for Exploring and Manipulating Networks", in *Proceedings of the Third International ICWSM Conference, San José, USA, 2009,* AAAI Press

Borges J. and Levene, M. (2006), "Ranking Pages by Topology and Popularity within Web Sites", *World Wide Web,* Vol. 9 No. 3, pp. 301-316

Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. (2003), "Model-Based Clustering and Visualization of Navigation Patterns on a Web Site", in *Data Mining and Knowledge Discovery,* Vol. 7 No. 4, pp. 399-424

Catledge, L. D. and Pitkow, J. E. (1995), "Characterizing browsing strategies in the World-Wide Web", *Comput. Networks and ISDN Systems,* Vol. 27 No. 6, pp. 1065-1073.

Chakrabarti, D., Kumar, R. and Punera, K. (2009), "Quicklink selection for navigational query results", in *Proceedings of the 18th international conference on World wide web, New York, USA, 2009,* ACM, pp. 391-400

Chen, Ch. (2004), Information Visualization, Spinger, 2nd edition,

Chen, J., Sun, L., Zaïane, O. R. and Goebel, R. (2004), "Visualizing and discovering web navigational patterns", in *Proceedings of the 7th International Workshop on the Web and Databases, Paris, France, 2004,* ACM, pp. 13-18

Chi, E. H. (2002), "Improving Web Usability through Visualization", *IEEE Internet Computing,* Vol. 6 No. 2, pp. 64-71

Chi, E. Pirolli, P. and Pitkow, J. (2000), " The scent of a site: A system for analyzing and predicting information scent, usage and usability of a web site", in *Proceedings of the ACM CHI 2000 Human Factors in Computing Systems Conference, The Hague, The Netherlands, 2000*, ACM, pp. 161-168

Cooley, R., Mobasher, B. and Srivastava, J. (1999), "Data Preparation for Mining World Wide Web Browsing Pattern", Knowledge and Information Systems, Vol. 1 No. 1, pp. 5-32

Cooley, R., Mobasher, B. and Srivastava, J. (1997), "Web Mining: Information and Pattern Discovery on the World Wide Web", in *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, USA, 1997,* IEEE Computer Society, pp. 558-567

da Cruz, S. M. S., Campos, M. L. M., Pires, P. F. and Campos, L. M. (2004), "Monitoring e-business Web services usage through a log based architecture Web Services", in *Proceedings of the IEEE International Conference on Web Services,* Washington, USA, 2004, IEEE Computer Society, pp. 61 - 69

Eirinaki, M. and Vazirgiannis, M. (2007), "Web site personalization based on link analysis and navigational patterns", *ACM Trans. Internet Technol.,* Vol. 7 No. 4, Article 21

Erbacher, R. B. (2001), "Visual Traffic Monitoring and Evaluation", in *Proceedings of the Conference on Internet Performance and Control of Network Systems II,* Denver, USA, 2001, pp. 153-160.

Erbacher, R. F. (2002), "Glyph-Based Generic Network Visualization", in *Proceedings of the SPIE '2002 Conference on Visualization and Data Analysis, San Jose, USA, 2002*, pp. 228-237

Forsati,R. and Meybodi, M. R. (2010), "Effective page recommendation algorithms based on distributed learning automata and weighted association rules", Expert Systems with Applications, Vol. 37 No. 2, pp. 1316-1330

Frécon, E. and Smith, G. (1998), "WebPath: A Three-Dimensional Web History," in *IEEE Symp. Information Visualization (Info-Vis 98), Piscataway, N.J., USA,* 1998, IEEE Press

Freeman, L. C. (1979), "Centrality in networks: I. conceptual clarification", Social Networks, Vol. 1, pp. 215-239.

Freeman, L. C. (1980), "The gatekeeper, pair-dependency, and structural centrality". Quality and Quantity, Vol. 14, pp. 585-592

Fry, B. (2004), *Computational Information Design*, [Ph. D. Thesis], MIT, Cambridge, USA

Gloor, P. A. and Zhoa, Y. (2004), "TeCFlow - A Temporal Communication Flow Visualizer for Social Network Analysis", *ACM CSCW Conference, Chicago, USA,* 2004

Lee, J., Hoch, R., Podlaseck, M., Schonberg, E. and Gomory, S. (2000), "Analysis and Visualization of Metrics for Online Merchandizing", in *Web Usage Analysis and User Profiling, International WEBKDD'99 Workshop, San Diego, California, USA, 1999*, Springer, pp. 126-141

He, D., Goker, A. and Harper, D. J. (2002), "Combining evidence for automatic Web session identification", Information Processing & Management, Vol. 38 No. 5, pp. 727-742

Herder, E. and Weinreich, H. (2005), "Interactive web usage mining with the navigation visualizer". In *CHI '05 extended abstracts on Human factors in computing systems (CHI EA '05), New York, NY, USA*, 2005, ACM, pp. 1451-1454

Hong, J. I. and Landay, J. A. (2001), "WebQuilt: A Framework for Capturing and Visualizing the Web Experience," in *10th International Conference on the World Wide Web,* 2001, Hong Kong, China, pp. 717-724

Lavene, M. (2005), *An Introduction to Search Engines and Web Navigation*, Pearson Education, London

Meiss, M. R., Menczer, F., Fortunato, S., Flammini, A. and Vespignani, A. (2008), "Ranking web sites with real user traffic", in *Proceedings of the international conference on Web search and web data mining* (WSDM '08), New York, NY, USA, ACM, pp. 65-76.

Minarik, P. and Dymacek, T. (2008), "NetFlow Data Visualization Based on Graphs", in *Proceedings of the 5th international workshop on Visualization for Computer Security* (VizSec '08), Cambridge, MA, USA, Springer-Verlag, pp. 144-151.

Munzner, T. and Burchard, P. (1995), "Visualizing the structure of the World Wide Web in 3D hyperbolic space", in *Proceedings of the first symposium on Virtual reality modeling language* (VRML '95), New York, NY, USA, CAM, pp. 33-38

Nanopoulos, A., Katsaros, D. and Manolopoulos, Y. (2001), "Effective Prediction of Web- User Accesses: A Data Mining Approach", in *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points*, San Francisco, CA, USA, 2001, Springer,

Nielsen, J. (1993), Usability Engineering, Morgan Kaufmann, San Francisco

W. de Nooy, A. Mrvar, V. Batagelj: *Exploratory Social Network Analysis with Pajek*, CUP, January 2005

Perkowitz, M. and Etzioni, O. (1999), "Adaptive web sites: automatically synthesizing web pages", in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, *MenloPark, CA, USA, 1998*, American Association for Artificial Intelligence, pp. 727–732

Perkowitz, M. and Etzioni, O. (1999), “Towards adaptive web sites: conceptual framework and case study”, in *Proceedings of the eighth international conference on World Wide Web*, *New York, NY, USA, 1999*, Elsevier North-Holland, pp. 1245–1258

Pitkow, J. and Bharat, K. A. (1994), “Webviz: a tool for world-wide web access log analysis”, in *Proceedings of the First International World-Wide Web Conference, Geneva, Switzerland, 1994*, pp. 271-277

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

Ting, I. H., Kimble, C. and Kudenko, D. (2004), "Visualizing and Classifying the Pattern of User's Browsing Behaviour for Website Design Recommendation", in *Proceedings of the First International Workshop on Knowledge Discovery in Data Stream, Pisa, Italy, 2004*, pp.101-102

Watts, D. J., and Strogatz, S. H. (1998), "Collective dynamics of 'small-world' networks", *Nature*, Vol. 393, pp. 440-442.

Yang, Q., Wang, H. and Zhang, W. (2002), “Web-log Mining for Quantitative Temporal-Event Prediction”, IEEE Computational Intelligence Bulletin, Vol. 1 No. 1

Youssefi, A. H., Duke, D. J. and Zaki, M. J. (2004), “Visual web mining”, in *Proceedings of the 13th international World Wide Web conference on Alternate track papers \& posters* (WWW Alt. '04), New York, USA, ACM, pp. 394-395

Zhu, J., Hong, J. and Hughes, J. G. (2004), “PageCluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation”. *ACM Transactions on Internet Technology*, Vol. 4 No. 2, pp. 185-208