Web site structure mining using social network analysis

M.R. Martínez-Torres
Escuela Universitaria de Estudios Empresariales, University of Seville, Seville, Spain
Sergio L. Toral
E.S. Ingenieros, University of Seville, Seville, Spain
Beatriz Palacios
Escuela Universitaria de Estudios Empresariales, University of Seville, Seville, Spain, and
Federico Barrero
E.S. Ingenieros, University of Seville, Seville, Spain

Abstract

Purpose – Web sites are typically designed attending to a variety of criteria. However, web site structure determines browsing behavior and way-finding results. The aim of this study is to identify the main profiles of web sites' organizational structure by modeling them as graphs and considering several social network analysis features.

Design/methodology/approach – A case study based on 80 institutional Spanish universities' web sites has been used for this purpose. For each root domain, two different networks have been considered: the first is the domain network, and the second is the page network. In both cases, several indicators related to social network analysis have been evaluated to characterize the web site structure. Factor analysis provides the statistical methodology to adequately extract the main web site profiles in terms of their internal structure.

Findings – This paper allows the categorization of web site design styles and provides general guidelines to assist designers to better identify areas for creating and improving institutional web sites. The findings of this study offer practical implications to web site designers for creating and maintaining an effective web presence, and for improving usability.

Research limitations/implications – The research is limited to 80 institutional Spanish universities' web sites. Other institutional university web sites from different countries can be analyzed, and the conclusions could be compared or enlarged.

Originality/value – This paper highlights the importance of the internal web sites structure, and their implications on usability and way-finding results. As a difference to previous research, the paper is focused on the comparison of internal structure of institutional web sites, rather than analyzing the web as a whole or the interrelations among web sites.

Keywords Web sites, Web site design, Inter-computer links, Social networking, Factor analysis, Spain

Paper type Research paper

The authors gratefully acknowledge support provided by the Spanish Ministry of Education and Science within the I + D + I national project with reference DPI2007-60128, and the Consejería de Innovación, Ciencia y Empresa (Research Project with reference P07-TIC-02621).

1. Introduction

The web is an enormous set of documents connected through hypertext links created by designers of web sites. Publishing on the web is more than just setting up a page on a site; it also usually involves linking to other pages on the web. The increasing amount of data available on the web provides a huge amount of useful information that can be processed to discover useful knowledge from the web (Roussinov and Zhao, 2003). This trend has conducted to "web mining" as a new emerging discipline. Broadly speaking, web mining can be defined as the discovery and analysis of useful information from the world wide web (Abedin and Sohrabi, 2009). It is a very active research field that involves the application of data mining techniques to the content, structure and usage of web resources. Although it derives from data mining, web mining has many unique characteristics (Fayyad *et al.*, 1996). For instance, the sources of web mining are web documents, which can be represented as a directed graph consisting of document nodes and hyperlinks. While the source of data mining is confined to the structural data in database, different kind of patterns can be identified in web mining considering the content of documents, the structure given by hyperlinks, or the way in which web pages are browsed (Jicheng *et al.*, 1999).

Basically, three areas of web mining are commonly distinguished, as shown in Figure 1: content mining, structure mining, and usage mining (Stumme *et al.*, 2006):

(1) Web content mining (WCM) deals with knowledge discovery in the web contents, including text, hypertext, images, audio and video. Recent advances in multimedia data mining promise to widen access also to image, sound, video, etc. content of web resources.

(2) Web structure mining (WSM) usually operates on the hyperlink structure of web pages. WSM focuses on sets of pages, ranging from a single web site to the web as a whole. WSM exploits the additional information that is (often implicitly) contained in the structure of hypertext. Therefore, an important application area is the identification of the relative relevance of different pages that appear equally pertinent when analyzed with respect to their content in isolation (Chakrabarti, 2003).

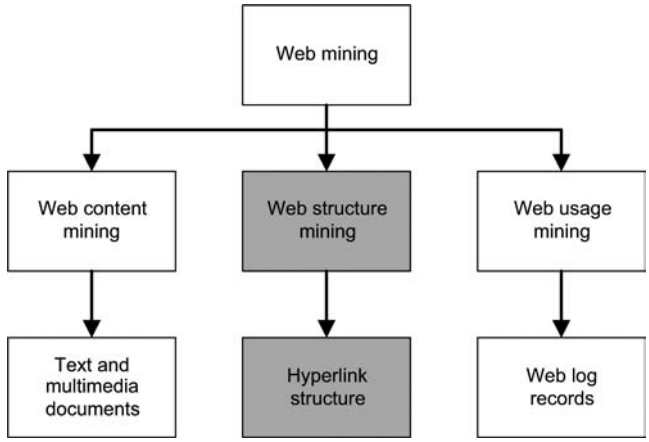

Figure 1.
Web mining categories and objects

(3) Web usage mining (WUM) focuses on records of the requests made by visitors to a web site, most often collected in a web server log (Arotaritei and Mitra, 2004). The content and structure of web pages, and in particular those of one web site, reflect the intentions of the authors and designers of the pages, and the underlying information architecture.

This paper is focused on WSM of institutional web sites. The study of web links can offer a valuable source of information, not only for developing informetric theory, but also for studying link patterns between network entities (Yang and Qin, 2008). In this case, universities' institutional web sites are studied representing them as graphs, and analyzing their features using social network analysis (SNA). An exploratory factor analysis is then performed to extract web site patterns according to their structure. The rest of the paper is organized as follows. The next section provides an overview about previous studies in the field of WSM, introducing SNA theory and its application to web site link analysis. The factor analysis methodology and the case study based on 80 Spanish universities are described in sections 3 and 4, respectively. In section 5, the proposed methodology is applied to the particular case study to extract the web site's structure patterns. Section 6 discusses the obtained results and their main implications, and finally, the conclusions are drawn.

2. WSM

The challenge for WSM is to deal with the structure of the hyperlinks within the web itself (Da Costa and Gong, 2005). The growing interest in web mining has led to a renewed interest on link analysis, which involves hypertext and web mining, relational learning and inductive logic programming, and graph mining (Getoor, 2003). Link structure evaluation and improvement is a significant problem which allows us to understand the overall web site structure and discover where information is concentrated or is missing (Abedin and Sohrabi, 2009). Several research lines can be distinguished when working with link analysis (Da Costa and Gong, 2005). Link-based classification deals with the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page. The Google's PageRank algorithm (Page *et al.*, 1998) is such an existing metric used for the evaluation of web pages. This algorithm is an important component of the Google's search engine, and boosted several studies on the link analysis of the web (Kleinberg, 1999; Snyder and Rosenbaum, 1999). A second group of studies are related to link-based cluster analysis, where data is segmented into groups, being similar objects grouped together, and dissimilar objects grouped into different groups. The final aim consists of discovering hidden patterns from data. In this case, the majority of studies are focused on the structure of the web considered in a large scale. For instance, the relationships among web domains have been analyzed in the Nordic academic web space (Ortega and Aguillo, 2008), or even in the world web space (Ortega and Aguillo, 2009).

Typically, SNA has been frequently used for the study of link analysis (Park and Thelwall, 2003; Toral *et al.*, 2010). SNA is a set of research procedures for identifying structures in social systems based on the relations among the system components, also referred to as nodes. In applying SNA methods to link analysis, web sites or web pages are considered the actors, and therefore the nodes in the social network graph, while

links are modeled as the relations between actors, represented by the edges of the graph (Iacobucci, 1994; Broder *et al.*, 2000). The resulting graph will be a directed graph because links are defined by an HTML tag within a markup file addressing a new web page setting the direction of the arc (in directed graphs, edges are called arcs).

Several studies making use of SNA can be found both for link-based classification and link-based cluster analysis studies. For the first group, SNA has been applied considering the Indegree method as an alternative to Pagerank methods. Link analysis through SNA has also been combined with text analysis to improve web information retrieval algorithms (Almpanidis *et al.*, 2007). For the second group, the relationships among web domains have been analyzed attending to several SNA criteria like degree and ranking (Baeza-Yates and Castillo, 2007; Ortega and Aguillo, 2009).

In this context, the purpose of this paper is to study web sites structure patterns by modeling web sites as connected graphs and by extracting several SNA features. Obtained results will highlight different web sites' profiles attending to their internal structure. This structure is closely related to users' navigation experience. Badly designed web sites frustrate users and cause them to leave as they cannot find what they need. The reasons cited for the users' negative experience include unavailability of information and, above all, difficulties for finding the required information. An adequate web site structure planning may improve accessibility and users' satisfaction (Woo and Kee, 2006).

Although web structure has frequently been studied, comparatively little is known at the web site level concerning its internal structure as an information organization and access mechanism. In this paper, web sites are modeled as two social networks. On the first network, nodes represent subdomains or external domains, and arcs represent the links among them. The second one is similar but considers web pages instead of domains or subdomains. A huge number of indicators related to different features of the derived networks can be computed using SNA.

A. SNA

SNA arose from using mathematical models of graphs applied in the analysis of social relationships between actors (Wasserman and Faust, 1994). In sociology, actors typically model individuals, groups, and occasionally autonomous devices. According to Wasserman and Faust (1994), "a social network consists of a finite set or sets of actors and the relation or relations defined on them". It is a complex system that is characterized by a high number of dynamically interconnected entities, and connects entities in any type of link that implies a peer-to-peer relationship (Bartal *et al.*, 2009). SNA may be viewed as a broadening or generalization of standard data analytic techniques and applied statistics that usually focus on observational items and their characteristics (Wasserman and Faust, 1994).

Although some alternatives based on content analysis have been proposed in the literature (Toral *et al.*, 2009a), Park and Thelwall (2003) states that, "compared to other web methods such as a content-based analysis, the relative advantage of hyperlink analysis is that it is able to examine the way in which web sites form a certain kind of relations with others via hyperlinks," and "using this information in combination with other web analyses can contribute to the understanding of why and how certain types of contents come to appear on web sites." Potgieter *et al.* (2007) indicates that SNA is a

research area aimed at understanding social complexity by representing and analyzing social networks using mathematical graphs.

Mathematically, a social network can be represented as a graph $G = (V, E)$ where V denotes a finite set of nodes and E denotes a finite set of arcs such that $E \subseteq V \times V$. Some network analysis methods are easier to understand when graphs are conceptualized as matrices (Nooy *et al.*, 2005), see equation (1):

$$M = \left(m_{ij}\right)_{n^*n} where\, n = |V|,\ m_{ij} = \begin{cases} 1\ if\left(v_i, v_j\right) \in E \\ 0 \quad otherwise \end{cases}. \tag{1}$$

In case of a valued graph, real valued weight function w(e) is defined on the set of arcs, i.e. $w(e) = E \rightarrow \Re$, and the matrix is then defined as given by equation (2):

$$m_{ij} = \begin{cases} w(e)\ if\left(v_i, v_j\right) \in E \\ 0 \qquad otherwise \end{cases}. \tag{2}$$

B. SNA features of web sites

Networks representing web sites are collected starting at a given page (the root of the institutional web site) and then following the out links to other pages. Two different kinds of networks are considered for each web site. The first one is the domain network in which nodes represent sub domains or external domains different to the root domain. Arcs represent the link among them. The second network is the page network containing all the web pages of the institutional web site and the links among them. Obviously, both networks are directed graphs, and they can be extracted to the desired depth. In both cases, network building is limited to the root domain. Although hyperlinks to other domains or pages outside the root domain are considered, their out links will not be followed.

In the context of link analysis, the referred domain network is a star network, with the root domain at the center of the star and the rest of domains linked with it. Several indicators related to the size of the domain network have been measured in terms of nodes and arcs. Typically, institutional web sites include sub domains, which should be distinguished from external domains. Therefore, this distinction has been made when considering the size in terms of nodes. Finally, the density and average degree of the network have also been considered as indicators. Density is related to the number of arcs, and degree is a measure of the number of links in which each node is involved (Martínez-Torres *et al.*, 2010).

The referred page network is a more complex network, with a higher size and a much higher number of arcs than the domain network. Consequently, a higher number of social network features can be extracted:

- *Size*. The number of nodes indicates the number of web pages and arcs represent the interrelations among these web pages. An important parameter to be chosen is the depth of link coverage when capturing web site information. A depth of seven has been used in this study. This value is considered sufficient to capture the essential information of web site structure, and is higher than the depth of five used in previous studies (Yang and Qin, 2008).

- *Density*. This is defined as the number of arcs in a simple network, expressed as a proportion of the maximum possible number of arcs. The main problem of this definition is that it does not take into account valued arcs higher than 1, and it depends on the network size. A different measure of density is based on the idea of the degree of a node, which is the number of arcs incident with it (Toral *et al.*, 2009b). A higher degree of nodes yields to a denser network, because nodes entertain more ties. Consequently, the average degree is a non-size dependent measure of density if multiple arcs and arc values are not disregarded. As the page network is a directed graph, several statistical measures of the out-degree distribution will be considered. Finally, density can be measured alternatively using an egocentric point of view; the egocentric density of a node is the density of ties among its neighbors (Nooy *et al.*, 2005).
- *Components*. A strong component is a maximal strongly connected subnetwork. A network is said to be strongly connected if each pair of nodes is connected by a path, taking into account the direction of arcs (Nooy *et al.*, 2005). In the context of this study, components allow the identification of connected substructures in the general web site.
- *K-cores*. A k-core is a sub-network in which each node has k degree in that sub-network. That means each node is connected to at least k other nodes. K-cores allow detecting groups with a strong link density that can reveal a latent structure in the data (Leydesdorff and Wagner, 2008). In the context of WSM, the core with the highest degree is the central core of the network, detecting the set of nodes where the network rests on. It has been used by Ortega and Aguillo (2008) to detect sub-networks among Nordic academic web sites.
- *Distance*. This is defined as the number of steps in the shortest path that connect two nodes. In the case of web sites, there is a clearly defined main node which is the root of the network. Consequently, it makes sense to measure the distance of pages to this node.
- *Closeness centralization*. This is an index of centrality based on the concept of distance. The closeness centrality of a node is calculated considering the total distance between one node and all other nodes, where larger distances yield lower closeness centrality scores. The closeness centralization is an index defined for the whole network, and it is calculated as the variation in the closeness centrality of nodes divided by the maximum variation in closeness centrality scores possible in a star network of the same size (Toral *et al.*, 2009c).
- *Betweenness*. This is a measure of centrality that rests on the idea that a person is more central if he or she is more important as an intermediary in the communication network (Nooy *et al.*, 2005). The centrality of a node depends on the extent to which this node is needed as a link to facilitate the connection of nodes within the network. Then, they are said to develop a brokerage role. If a geodesic is defined as the shortest path between two nodes, the betweenness centrality of a node is the proportion of all geodesics between pairs of other nodes that include this node, and betweenness centralization of the network is the variation in the betweenness centrality of nodes divided by the maximum variation in betweenness centrality scores possible in a network of the same size.

From the link analysis perspective, this measure allows detecting gateways connecting separate sub networks (Faba-Pérez *et al.*, 2005).

- *Partition correlation*. A partition of a network is a classification or clustering of the nodes in the network such that each node is assigned to exactly one class or cluster (Toral *et al.*, 2010). Two important partitions can be extracted using network features previously introduced. The first one is the k-neighbor partition, in which nodes are clustered using the distance to the root node. The second one is the out-degree partition in which nodes are clustered attending to their out-degree value. The correlation between both partitions is related to the extent in which the web site is following a tree structure from the root domain. Two types of association indices are computed: Cramer's V and Rajski's information index (Nooy *et al.*, 2005). Cramer's V is a statistic which measures the strength of association or dependency between two nominal or categorical variables. It is derived directly from the Chi-square statistic computed for a given crosstabulation or contingency table, and its value is reported in the range [0, 1]. The closer Cramer's V is to 0, the smaller association between the two variables. Rajski's indices measure the degree to which the information in one classification is preserved in the other classification. Given two classifications $A = (A_1, A_2, \ldots, A_k)$ and $B = (B_1, B_2, \ldots, B_m)$ from a dataset X, the Rajski coefficient is calculated as $R(A,B) = \sqrt{1 - d^2(A,B)}$, where d(A, B) is a measure of the mutual information held in common between the subsets of two classifications. The value of R(A; B) varies between zero and unity, indicating respectively the degree of relatedness from none to perfect (Orlozi, 1968).

3. Methodology

Factor analysis has been applied to categorize web sites according to the style in which they have been designed. Factor analysis is a data reduction technique used to find homogeneous groups in a large set of data. It addresses the problem of analyzing the structure of interrelationships among a number of variables by defining a set of common underlying dimensions (Hair *et al.*, 1995). These groups represent the underlying variables or factors, which can explain the pattern of correlations within a set of observed variables (Stevens, 1992). Each observable variable is assumed to be dependent on a linear combination of the common factors, and the coefficients are known as loadings (Rencher, 2002). There are several extraction methods: principal components and principal axis factoring (or principal factor analysis) are among the most widely used. According to Hair *et al.* (1995), the former is used when the objective is to summarize most of the original information in a minimum number of factors, whereas the latter is used to identify the underlying dimensions reflecting what the variables share in common. In most applications both methods arrive at essentially identical results.

Factor analysis can be used for either exploratory or confirmatory purposes: exploratory analyses do not set any a priori constraints on the estimation of factors or the number of factors to be extracted, while confirmatory analysis does. In our case, we have developed an exploratory analysis, as we did not know the number of underlying dimensions. That means a decision must be made about the number of factors to be extracted. There are several criteria for doing this, being the most extensive the eigenvalue and percentage of variance criterion. The percentage of variance criterion

considers all factors accounting for about 70 percent of the variance of the original variables (Hair *et al.*, 1995).

Once the number of factors has been determined, the next step is to interpret them according to the factor loadings matrix. The estimated loadings from an unrotated factor analysis fit can usually have a complicated structure. The goal of orthogonal factor rotation is to find a parameterization in which each variable has only a small number of large loadings, i.e. is affected by a small number of factors. The rotated factor analysis fit ensures that factors represent unidimensional constructs while preserving the essential properties of the original loadings. The most popular of these techniques is the varimax rotation, which seeks rotated loadings that maximize the variance of the squared loadings in each column of the factor loading matrix (Rencher, 2002).

Factor scores, which are defined as estimates of the underlying factor values for each observation, can also be obtained from factor analysis. Using factor scores, each of the observations can be assigned to one or none of the latent factors. The assignation has been performed using the maximum factor score, provided that this maximum value is higher than 0.1 (Rencher, 2002). Factor scores can be used as inputs for other statistical analyses, like ANOVA, to check the categorization of the original sample leads to significant different groups.

4. Case study

The case study includes up to 80 Spanish university web sites. All of them are included in the webometrics Ranking of World Universities (www.webometrics.org), where more than 6,000 universities all over the world are sorted according to size and visibility. Table I lists the root domains of the considered web sites.

They cover almost the whole range of webometrics ranking, and exhibit a wide variety of sizes in terms of domains and web pages. More than 718,000 web pages and more than four million out links have been considered through the analysis of this web sites list. For each web site, two starting networks have been collected: the domain network and the page network. As an example, Figures 2 and 3 show the particular case of the domain and page network, respectively, corresponding to the University of Seville web site.

The social network features previously described have been measured, considering sometimes the whole network and sometimes the subnetworks excluding nodes with 0 out-degree or subnetworks with $k > 1$ cores. As a result, 24 indicators have been obtained (Table II).

5. Results

Factor analysis has been applied to the data extracted from the mentioned group of universities' web sites. The eigenvalues of the sample covariance matrix are shown in Table III. The number of factors able to account for more than 70 percent of the total sample variance is four. As it will be demonstrated later using the analysis of variance, this number of factors also leads to a categorization of the original data sample into significant different groups

The resulting aggregation of variables leads to the identified latent factors of Table IV.

Spanish universities' web sites			
www.ucm.es/	http://portal.uned.es/	www.ual.es/	www.cef.es/
www.upc.edu/	www.uva.es/	www.udl.es/	www.uch.ceu.es/
www.upm.es/	www.upf.edu/	www.ujaen.es/	www.nebrija.com/
www.uab.es/	www.unav.es/	www.umh.es/	www.uic.es/
www.ehu.es/	www.uc3m.es/	www.deusto.es/	www.url.es/
www.ub.edu/	www.uniovi.es/	www.unavarra.es/	www.esdi.es/
www.us.es/	www.uma.es/	www.upct.es/	www.uax.es/
www.upv.es/	www.uco.es/	www.upo.es/	www.vives.org/
www.um.es/	www.ull.es/	www.ie.edu/	www.uimp.es/
www.ugr.es/	www.udc.es/	www.upcomillas.es/	www.ucjc.edu/
www.ua.es/	www.unex.es/	www.ceu.es/	https://www.ucv.es/
www.uvigo.es/	www.uah.es/	www.iese.edu/	www.uspceu.com/
www.uv.es/	www.uoc.edu/	www.ubu.es/	www.cesdonbosco.com/
www.uam.es/	www.udg.edu/	www.urv.net/	www.ufv.es/
www.usal.es/	www.ulpgc.es/	www.unirioja.es/	www.esic.es/
www.uji.es/	www.unican.es/	www.uem.es/	www.cepade.es/
www.unizar.es/	www.unileon.es/	www.esade.edu/	www.eoi.es/portal/
www.usc.es/	www.urjc.es/	www.ucam.edu/	www.esmuc.net/
www.uib.es/ca/	www.uca.es/	www.mondragon.edu/	www.udima.es/
www.uclm.es/	www.uhu.es/	www.uvic.es/	www.eupmt.es/

Table I.
List of considered web sites

Using factor scores, the original sample of universities can be approximated to one of the identified latent factors. Consequently, the original population of institutional universities' web sites can be split in four different populations associated to each factor. An analysis of variance (ANOVA) has been performed to check the null hypothesis of equal population means (Martínez-Torres and Toral, 2010). Table V shows the F statistic, the ratio of two different estimators of population variance, which appears together with its corresponding critical level or observed significance. Results of Table V mean that the null hypotheses can be rejected in all the cases with a significance value below 0.05.

Using the classification given by factor scores, the mean values of the selected indicators for each factor are calculated (Table VI). The identified factors of Table IV and the mean value of indicators per factor of Table VI have been used to distinguish the following web sites structure patterns:

(1) Factor 1 represents highly structured web sites. The high value of Rajski and Cramer's V information indices indicates that out-degree grows as nodes are more distant from the root domain. The high value of average value and the standard deviation of nodes betweenness centrality suggest the web site is structured through highly interconnected nodes spread over the web site, following a certain tree structure. Finally, factor 1 exhibits a high value of density due to the fact of being small web sites as compared to the web sites assigned to other factors.

Figure 4 is a symbolic representation of web sites identified by factor 1. It is a highly structured web site, where information finding requires browsing through several web pages. If node B represents the webpage with the required

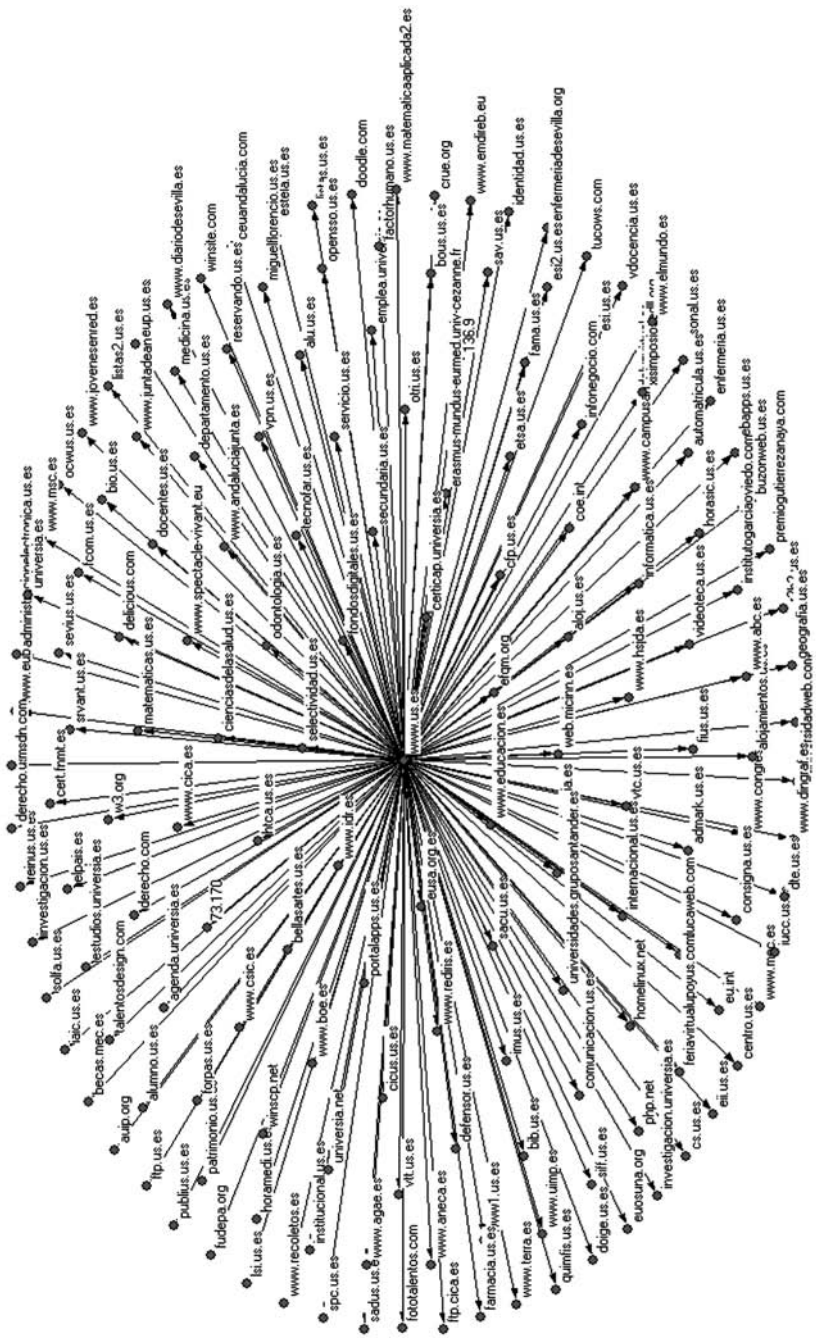

Figure 2.
University of Seville domain network

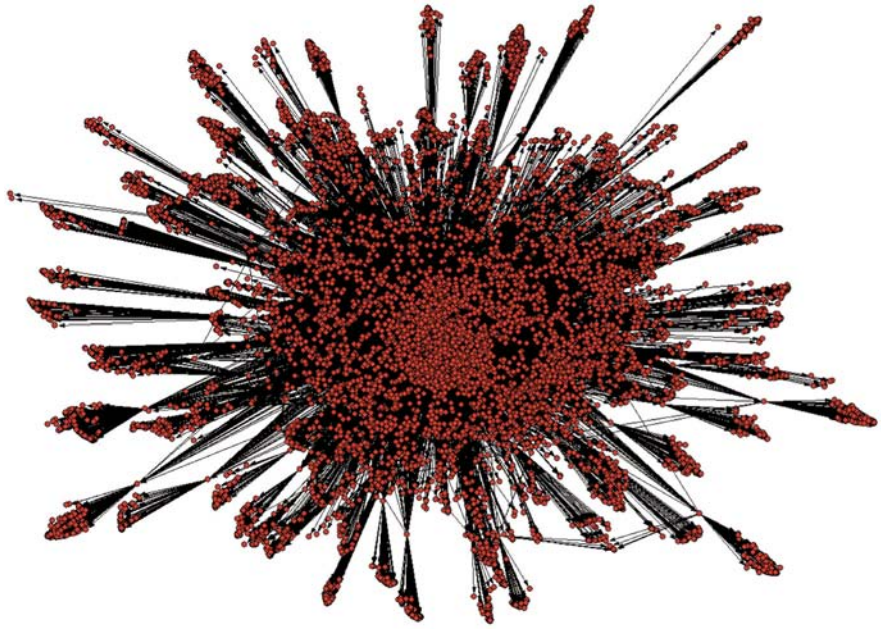

Figure 3.
University of Seville page network

information, several mouse clicks are necessary to reach the desired information.

(2) Factor 2 represents large web sites, which probably have been growing during the years in a certain chaotic progression. The number of pages grows geometrically with the depth level, so a long navigation process is necessary to achieve the desired information. Most of the web pages play a betweenness role, as there is not a formal structure under which the web site was designed. The high mean value of indicators associated to factor 2 (I2, I3, I5, I8 and I13) support this conclusion.

Figure 5 shows the non-structured organization of this kind of web site. Accessibility of information may be confusing if the user is not a regular user of the web site.

(3) Factor 3 represents partitioned web sites, where the global network could be considered as the sum of more or less independent subnetworks. In this case, web sites are organized around subdomains related to different areas of the organization. The high out-degree and egocentric values justify this kind of organization.

Figure 6 represents the interconnected subnetworks available from the root domain A.

(4) Finally, factor 4 represents a more centralized structure in the sense of distance to the root domain. There is a core of highly interconnected pages around the root domain, facilitating the accessibility of information. The flat structure represented in Figure 7 means a short distance between the root domain A and the required information B.

	Description	Network
I1	Density (domain network)	Domain network
I2	Number of pages	Page network
I3	Total number of arcs	Page network
I4	Density (page network)	Page network
I5	Number of pages in the last level	Page network
I6	Average out-degree	Page network
I7	Standard deviation of out-degree	Page network
I8	Number of pages	Page network (excluding nodes with out-degree = 0)
I9	Density (page network excluding out-degree = 0)	Page network (excluding nodes with out-degree = 0)
I10	Average degree	Page network (excluding nodes with out-degree = 0)
I11	Average out-degree	Page Network (excluding nodes with out-degree = 0)
I12	Standard deviation of closeness centrality	Page Network (excluding nodes with out-degree = 0)
I13	Number of nodes with betweenness centrality > 0	Page network
I14	Standard deviation of nodes betweenness centrality	Page network of k-cores, k > 0
I15	Average value of nodes betweenness centrality	Page network (excluding nodes with out-degree = 0)
I16	Standard deviation of nodes betweenness centrality	Page network (excluding nodes with out-degree = 0)
I17	Rajski(C1 ↔ C2)	Page network
I18	Rajski(C1 ← C2)	Page network (excluding nodes with out-degree = 0)
I19	% of pages included in strong components	Page network
I20	Average value of closeness centrality	Page network
I21	Standard deviation of closeness centrality	Page network
I22	Cramer's V	Page network (excluding nodes with out-degree = 0)
I23	Egocentric density (average value)	Page network (excluding nodes with out-degree = 0)
I24	Egocentric density (average value)	Page network of k-cores, k > 0

Table II.
List of selected indicators

Basically, identified profiles of web site structures respond to two basic strategies when deciding their final structure (Tan and Wei, 2006). The first strategy consists of offering a structure that makes sense to the final user. In this sense, web sites sacrifice accessibility of information looking for a more structured navigation scheme. Factors 1, 2 and 3 could be included in this strategy. The alternative option consists of reducing big structures under the assumption that user performance is optimal when breadth and depth of web site is kept to a moderate level (Tan and Wei, 2006). This is the strategy represented by factor 4.

6. Discussion and implications

Prior studies consider four different navigation structure types: a tree, a tree with a return-to-home page button, a tree with a few horizontal links, and an extensive

Factor	Eigenvalues Total	% of variance	Cumulative %
1	10.389	33.511	33.511
2	5.557	17.924	51.436
3	3.443	11.108	62.543
4	2.097	6.764	69.307
5	1.294	4.175	73.482
6	1.287	4.152	77.634
7	1.029	3.321	80.955
8	0.942	3.038	83.993
...	...	...	...
...	...	...	...
30	0.005	0.018	99.994
31	0.002	0.006	100.000

Table III.
Total variance explained

		Factor loading
Factor 1: highly structured web sites		
I1	Density (domain network)	0.731
I4	Density (page network)	0.617
I9	Density (page network excluding out-degree = 0)	0.728
I12	Standard deviation of closeness centrality	0.659
I14	Standard deviation of nodes betweenness centrality	0.840
I15	Average value of nodes betweenness centrality	0.808
I16	Standard deviation of nodes betweenness centrality	0.800
I17	Rajski(C1 $\leftrightarrow$ C2)	0.716
I18	Rajski(C1 $\leftarrow$ C2)	0.750
I22	Cramer's V	0.683
Factor 2: large web sites		
I2	Number of pages	0.885
I3	Total number of arcs	0.912
I5	Number of pages in the last level	0.779
I8	Number of pages	0.841
I13	Number of nodes with betweenness centrality > 0	0.873
Factor 3: partitioned web sites		
I6	Average out-degree	0.726
I7	Standard deviation out-degree	0.624
I10	Average degree	0.878
I11	Average out-degree	0.715
I24	Egocentric density (average value)	0.687
I23	Egocentric density (average value)	0.713
Factor 4: centralized web sites		
I19	% of pages included in strong components	0.748
I20	Average value of closeness centrality	0.930
I21	Standard deviation of closeness centrality	0.785

Table IV.
Identified factors

Notes: Extraction method: Principal Component Analysis; Rotation method: VARIMAX with Kaiser normalization; the rotation has converged in six iterations

	F	Sig.
I1	6.472	0.000
I2	10.589	0.000
I3	18.079	0.000
I4	4.509	0.003
I5	2.95	0.025
I6	11.767	0.000
I7	7.896	0.000
I8	38.082	0.000
I9	6.225	0.000
I10	26.949	0.000
I11	13.405	0.000
I12	6.837	0.000
I13	33.305	0.000
I14	28.987	0.000
I15	27.906	0.000
I16	30.786	0.000
I17	12.002	0.000
I18	10.797	0.000
I19	15.825	0.000
I20	13.567	0.000
I21	13.58	0.000
I22	11.655	0.000
I23	9.214	0.000
I24	7.682	0.000

Table V.
Statistical significance of ANOVA

	F1	F2	F3	F4
I1	0.025	0.003	0.003	0.015
I2	1,329.667	28,783.538	9,145.118	2,236.833
I3	5,629.667	147,043.462	94,650.529	19,714.250
I4	0.014	0.000	0.002	0.010
I5	240.000	12,834.000	940.176	85.667
I6	4.942	6.463	11.793	9.454
I7	13.925	19.526	26.024	17.312
I8	190.800	3,884.462	1,496.000	771.167
I9	0.106	0.009	0.055	0.047
I10	26.662	52.124	98.715	43.537
I11	25.708	37.435	57.524	25.541
I12	0.106	0.055	0.052	0.058
I13	162.467	3,570.923	1,303.412	617.417
I14	0.047	0.011	0.017	0.023
I15	0.010	0.001	0.002	0.004
I16	0.043	0.010	0.016	0.023
I17	0.517	0.321	0.352	0.441
I18	0.274	0.074	0.130	0.151
I19	8.558	4.173	11.227	32.370
I20	0.059	0.035	0.048	0.122
I21	0.111	0.076	0.097	0.140
I22	0.705	0.389	0.547	0.550
I23	0.250	0.203	0.305	0.303
I24	0.531	0.446	0.653	0.601

Table VI.
Mean values of selected indicators

network (Huizingh, 2000). Several of their proposed navigation structure types have also been identified in this study. For instance, the tree structure is related to the identified depth tree structure (factor 1), flat-tree structure (factor 4) and partitioned tree structure (factor 3), while extensive network type is related to identified large web sites (factor 2).

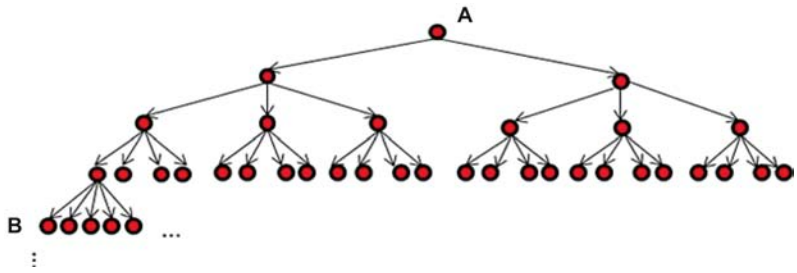

Figure 4.
Symbolic representation of factor 1 web sites

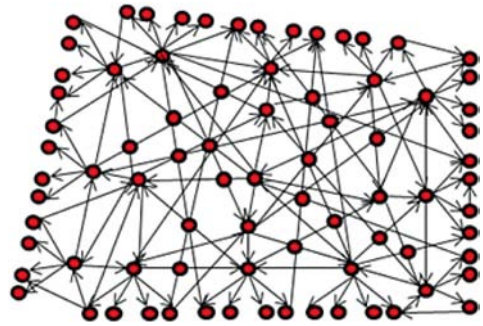

Figure 5.
Symbolic representation of factor 2 web sites

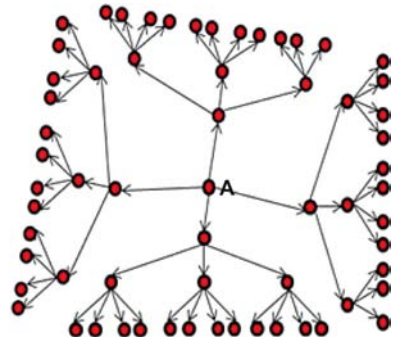

Figure 6.
Symbolic representation of factor 3 web sites

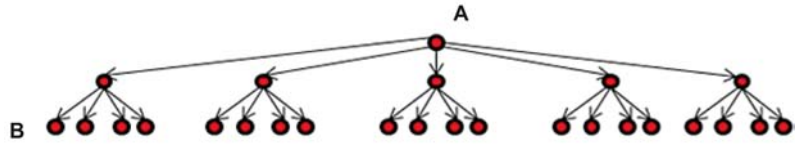

Figure 7.
Symbolic representation of factor 4 web sites

Web site internal structure is strongly related to issues like accessibility and navigability through web sites. Navigation features allow the visitor to get easy access to information of interest, both internal and external to the site. It is included as one of the design features of corporate web sites, along with presentation, security, speed and tracking (Robbins and Stylianou, 2003). The quality of a web site is also increased if the site is easily identifiable and accessible to the users. In fact, accessibility is part of web assessment indexes (Miranda González and Bañegil, 2004).

The web site organization has also important implications on usability. This refers to "how well and how easily a user, without formal training, can interact with an information system or web site" (Benbunan-Fich, 2001; Bar-Ilan, 2005). Better structure of web links enables visitors to navigate through web sites more easily, and also gets them to the right place sooner. Although web site usability can be evaluated following different approaches like cognitive walkthrough, Markov chains or survey methods (Abedin and Sohrabi, 2009), all of them are conditioned by the way the web site is structured. Consequently, the identified web site structure patterns should be considered by web site designers to improve usability.

Another important implication of this study refers to search engine optimization (SEO). The significant majority of online travel searches utilize a search engine as the initial point of entry. The three key methods for increasing visibility of web sites are SEO, search engine advertising, and paid inclusion. SEO involves adopting methods that improve the ranking of a web site when a user types in relevant keywords in a search engine; search engine advertising refers to buying display positions at the paid listing area of a search engine; and paid inclusion refers to paying search engine companies for the inclusion of the site in their organic listings. SEO is generally recognized the most effective one, as searchers pay less attention to commercial content than they do to organic listings (Jansen and Resnick, 2005). The structure of a site has a direct bearing on how well it can be perceived by search engines. Search engines send programs, known variously as "robots, "spiders" or "crawlers", to investigate the Internet and to find out what is on sites. Not just this, they use algorithms to process the data returned to them by crawlers, and to determine the relevance and popularity of sites to be listed on their resulting pages. Site structure is important because it will always affect the ways in which "crawlers" see it and its content. SEO services agree that if there are three or more links to each and every page from others on the site, then it can be said to be well structured for search engine optimization. If the structure of a site inhibits the number of internal links, then it is not well structured. Crawlers should be able to read as many pages as possible, and these links are the paths they will have to take. A site needs to be structured in such a way to allow easy and readily available navigation to and from a site map. A user-friendly structure of a site can generally be said to work well in terms of crawlers. That means search engine results page ranking can be improved with good site structure. The result page on a search engine gives the user their first glimpse of content after they have entered the text into the query box and hit enter. Web site structure can be seen as one of the SEO tools at web designers' disposal for making web site visible on a search engine results page.

7. Conclusion

This paper proposes the identification of web structure patterns using SNA techniques. As a case study, SNA features from 80 institutional web sites corresponding to Spanish

universities have been extracted and statistically analyzed. Results identify four types of web sites organization according to their structure. Three of them show different kinds of structured organization while the last one is closer to a flat organization, emphasizing the accessibility of information. Obtained results offer interesting implication for web designers. In particular, web site designers can benefit from this research by receiving tools for improving web site usability, which in turn benefits web site visitors. They can also check whether the structure of their web site is as they intended it to be. Finally, web sites should be structured for facilitating search engines to browse their contents. Two limitations can be mentioned associated to this study. The first one is that the study is restricted to Spanish universities, although it could be extended to universities all over the world or even to different institutional web sites as a future work. The second limitation refers to the number of used indicators. Although 24 indicators have been obtained from the dataset, some other SNA features, or the same features applied to different subnetworks, could be measured. A confirmatory work using a higher number of indicators could also be a natural extension of this work.

References

Abedin, B. and Sohrabi, B. (2009), "Graph theory application and web page ranking for web site link structure improvement", *Behaviour & Information Technology*, Vol. 28 No. 1, pp. 63-72.

Almpanidis, G., Kotropoulo, C. and Pitas, I. (2007), "Combining text and link analysis for focused crawling – an application for vertical search engines", *Information Systems*, Vol. 32 No. 6, pp. 886-908.

Arotaritei, D. and Mitra, S. (2004), "Web mining: a survey in the fuzzy framework", *Fuzzy Sets and Systems*, Vol. 148 No. 1, pp. 5-19.

Baeza-Yates, R. and Castillo, C. (2007), "Characterization of national web domains", *ACM Transactions on Internet Technology*, Vol. 7 No. 2, pp. 1-32.

Bar-Ilan, J. (2005), "What do we know about links and linking? A framework for studying links in academic environments", *Information Processing and Management*, Vol. 41 No. 4, pp. 973-86.

Bartal, A., Sasson, E. and Ravid, G. (2009), "Predicting links in social networks using text mining and SNA", *International Conference on Advances in Social Network Analysis and Mining, 2009, ASONAM 09*, pp. 131-6.

Benbunan-Fich, R. (2001), "Using protocol analysis to evaluate the usability of a commercial web site", *Information and Management*, Vol. 39 No. 2, pp. 151-63.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000), "Graph structure in the web", *Computer Networks and ISDN Systems*, Vol. 33 Nos 1-6, pp. 309-20, available at: http://www9.org/w9cdrom/160/160.html (accessed July 1, 2002).

Chakrabarti, S. (2003), *Mining the Web*, Morgan Kaufmann, San Francisco, CA.

Da Costa, M.G. and Gong, Z. (2005), "Web structure mining: an introduction", *IEEE International Conference on Information Acquisition*, pp. 590-5.

Faba-Pérez, C., Zapico-Alonso, F., Guerrero-Bote, V.P. and de Moya-Anegón, F. (2005), "Comparative analysis of webometric measurements in thematic environments", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 8, pp. 779-85.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, Vol. 39 No. 11, pp. 27-34.

Getoor, L. (2003), "Link mining: a new data mining challenge", *ACM SIGKDD Explorations Newsletter*, Vol. 4 No. 2, pp. 84-9.

Hair, J.F. Jr, Anderson, R.E., Tatham, R.L. and Black, W.C. (1995), *Multivariate Data Analysis with Readings*, Prentice Hall International, London.

Huizingh, E.K. (2000), "The content and design of web sites: an empirical study", *Information & Management*, Vol. 37 No. 3, pp. 123-34.

Iacobucci, D. (1994), "Graphs and matrices", in Wasserman, S. and Faust, K. (Eds), *Social Network Analysis – Methods and Applications*, Cambridge University Press, New York, NY, pp. 92-166.

Jansen, B.J. and Resnick, M. (2005), "Examining searcher perceptions of and interactions with sponsored results", paper presented at the Workshop on Sponsored Search Auctions at ACM Conference on Electronic Commerce (EC05), Vancouver, June 5-8.

Jicheng, W., Yuan, H., Gangshan, W. and Fuyan, Z. (1999), "Web mining: knowledge discovery on the web", *IEEE International Conference on Systems, Man, and Cybernetics SMC 99*, Vol. 2, pp. 137-41.

Kleinberg, J. (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46 No. 5, pp. 604-32.

Leydesdorff, L. and Wagner, C.S. (2008), "International collaboration in science and the formation of a core group", *Journal of Informetrics*, Vol. 2 No. 4, pp. 317-25.

Martínez-Torres, M.R. and Toral, S.L. (2010), "Strategic group identification using evolutionary computation", *Experts Systems with Applications*, Vol. 37 No. 7, pp. 4948-54.

Martínez-Torres, M.R., Barrero, F., Cortés, F. and Toral, S.L. (2010), "The role of Internet in the development of future software projects", *Internet Research*, Vol. 20 No. 1, pp. 72-86.

Miranda González, F.J. and Bañegil, T.M. (2004), "Quantitative evaluation of commercial web sites: an empirical study of Spanish firms", *International Journal of Information Management*, Vol. 24 No. 4, pp. 313-28.

Nooy, W., Mrvar, A. and Batagelj, V. (2005), *Exploratory Network Analysis with Pajek*, Cambridge University Press, Cambridge.

Orlozi, L. (1968), "Information analysis in phytosociology: partition, classification and prediction", *Journal of Theoretical Biology*, Vol. 20 No. 3, pp. 271-84.

Ortega, J.L. and Aguillo, I.F. (2008), "Visualization of the Nordic academic web: link analysis using social network tools", *Information Processing and Management*, Vol. 44 No. 4, pp. 1624-33.

Ortega, J.L. and Aguillo, I.F. (2009), "Mapping world-class universities on the web", *Information Processing and Management*, Vol. 45 No. 2, pp. 272-9.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), "The PageRank citation ranking: bringing order to the web", available at: http://dbpubs.stanford.edu/pub/1999-66 (accessed May 22, 2009).

Park, H. and Thelwall, M. (2003), "Hyperlink analysis of the world wide web: a review", *Journal of Computer Mediated Communication*, Vol. 8 No. 4, available at: http://jcmc.indiana.edu/vol8/issue4/park.html (accessed January 28, 2009).

Potgieter, A., April, K.A., Cooke, R.J.E. and Osunmakinde, I.O. (2007), "Temporality in link prediction: understanding social complexity", *Sprouts: Working Papers on Information Systems*, Vol. 7 No. 9.

Rencher, A.C. (2002), *Methods of Multivariate Analysis*, 2nd ed., Wiley Series in Probability and Statistics, Wiley & Sons, New York, NY.

Robbins, S.S. and Stylianou, A.C. (2003), "Global corporate web sites: an empirical investigation of content and design", *Information & Management*, Vol. 40 No. 3, pp. 205-12.

Roussinov, D. and Zhao, J.L. (2003), "Automatic discovery of similarity relationships through web mining", *Decision Support System*, Vol. 35 No. 1, pp. 149-66.

Snyder, H. and Rosenbaum, H. (1999), "Can search engines be used for web-link analysis? A critical review", *Journal of Documentation*, Vol. 55 No. 4, pp. 375-84.

Stevens, J. (1992), *Applied Multivariate Statistics for the Social Sciences*, 2nd ed., Lawrence Erlbaum, Mahwah, NJ.

Stumme, G., Hotho, A. and Berent, B. (2006), "Semantic web mining: state of the art and future directions", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 4 No. 2, pp. 124-43.

Tan, G.W. and Wei, K.K. (2006), "An empirical study of web browsing behaviour: towards an effective web site design", *Electronic Commerce Research and Applications*, Vol. 5 No. 4, pp. 261-71.

Toral, S.L., Barrero, F. and Martínez-Torres, M.R. (2009a), "Knowledge sharing through online communities of practice: the case of linux ports to embedded processors", *Proceedings of the IADIS International Conference on Web Based Communities (IADIS-09)*, pp. 107-12.

Toral, S.L., Martínez-Torres, M.R. and Barrero, F. (2009b), "Virtual communities as a resource for the development of OSS projects: the case of Linux ports to embedded processors", *Behavior and Information Technology*, Vol. 28 No. 5, pp. 405-19.

Toral, S.L., Martínez-Torres, M.R., Barrero, F. and Cortés, F. (2009c), "An empirical study of the driving forces behind online communities", *Internet Research*, Vol. 19 No. 4, pp. 378-92.

Toral, S.L., Martínez-Torres, M.R. and Barrero, F. (2010), "Analysis of virtual communities supporting OSS projects using social network analysis", *Information and Software Technology*, Vol. 52 No. 3, pp. 296-303.

Wasserman, S. and Faust, K. (Eds) (1994), *Social Network Analysis – Methods and Applications*, Cambridge University Press, New York, NY.

Woo, G. and Kee, K. (2006), "An empirical study of web browsing behaviour: towards an effective web site design", *Electronic Commerce Research and Applications*, Vol. 5 No. 4, pp. 261-71.

Yang, B. and Qin, J. (2008), "Data collection system for link analysis", *Third International Conference on Digital Information Management, ICDIM 2008*, pp. 247-52.

Further reading

Berlt, K., Silva de Moura, E., Carvalho, A., Cristo, M., Ziviani, N. and Couto, T. (2010), "Modeling the web as a hypergraph to compute page reputation", *Information Systems*, Vol. 35 No. 5, pp. 530-43.

Dodge, M. (n.d.), "Cyber-geography research", available at: www.cybergeography.org/home.html (accessed April 20, 2007).

Kim, H.J. (2000), "Motivations for hyperlinking in scholarly electronic articles: a qualitative study", *Journal of the American Society of Information Science and Technology*, Vol. 51 No. 10, pp. 887-99.

Paterson, R. and Cox, D. (n.d.), "Visualization study of the Nsfnet", available at: http://vis.ncsa.uiuc.edu/?content=projects&subcontent=show&ID=4 (accessed April 20, 2007).

Thelwall, M. (2003), "Can Google's PageRank be used to find the most important academic web pages?", *Journal of Documentation*, Vol. 59 No. 2, pp. 205-17.

Zook, M.A. (2000), "The web of production: the economic geography of commercial Internet content production in the United States", *Environment and Planning A*, Vol. 32, pp. 411-26.

About the authors

M.R. Martínez-Torres is an Associate Professor in Management and Business Administration at Business Administration and Marketing Department, University of Seville. Her main research interests include Intellectual Capital and Knowledge Management, and Social Network Analysis. She has co-authored articles in many leading academic and professional journals, including: *Information and Management*, *IEEE Transactions on Education*, *Computers & Education*, and *Behaviour and Information Technology*.

Sergio L. Toral is an Associate Professor in Digital Electronic Systems at the Department of Electronic Engineering, University of Seville. His main research interests include microprocessor and DSP devices, real-time systems, open source software projects and social network analysis. Sergio L. Toral is the corresponding author and can be contacted at: toral@esi.us.es

Beatriz Palacios is Doctorate Student at Business Administration and Marketing Department, University of Seville. Her main research interests include link analysis and webometrics.

Federico Barrero is an Associate Professor in Digital Electronic Systems at the Department of Electronic Engineering, University of Seville. His main research interests include microprocessor and DSP devices, embedded systems and their industrial applications.