An evolutionary factor analysis computation for mining website structures

M.R. Martínez-Torres [a,1], S.L. Toral [b,*], B. Palacios [a,1], F. Barrero [b]

[a] *Facultad de Turismo y Finanzas, University of Seville, Avda. San Francisco Javier s/n, 41018 Sevilla, Spain*
[b] *Escuela Superior de Ingenieros, University of Seville, Avda. Camino de los Descubrimientos s/n, 41092 Sevilla, Spain*

ARTICLE INFO

Keywords:
Link analysis
Website structure
Factor analysis
Evolutionary computation
Genetic algorithms

ABSTRACT

This paper explores website link structure considering websites as interconnected graphs and analyzing their features as a social network. Two networks have been extracted for representing websites: a domain network containing subdomains or external domains linked through the website and a page network containing webpages browsed from the root domain. Factor analysis provides the statistical methodology to adequately extract the main website profiles in terms of their internal structure. However, due to the large number of indicators, the task of selecting a representative subset of indicators becomes unaffordable. A genetic search of an optimum subset of indicators is proposed in this paper, selecting a multi-objective fitness function based on factor analysis results. The optimum solution provides a coherent and relevant categorization of website profiles, and highlights the possibilities of genetic algorithms as a tool for discovering new knowledge in the field of web mining.

1. Introduction

Link analysis is the quantitative study of hyperlinks between web pages. It is usually included as part of webometrics, which is the quantitative analysis of web phenomena, dealing also with web citation analysis, search engine evaluation and purely descriptive studies of the web (Björneborn & Ingwersen, 2004; Thelwall, 2008). Web links have been heavily studied during the last years in order to understand the structure and growth patterns of the Web (Thelwall, 2004), and they have been applied to the development of page ranking algorithms. The rapid development experienced by Web links analysis in the theories, technologies, and methodologies can be explained by the fact of being studied from different points of views, like computer science, information science, communications studies and sociology (Thelwall, 2004).

Social network analysis (SNA) has been frequently used for the study of link analysis (Park & Thelwall, 2003; Toral, Martínez Torres, & Barrero, 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, websites or web-pages are considered the actors, representing the nodes in the social network graph, while links are modeled as the relations between actors, representing the edges of the graph (Iacobucci, 1994). The resulting graph will be a directed graph because links are defined by an HTML tag within a markup file which address to a new web page setting the direction of the arc (in directed graphs, edges are called arcs).

The majority of studies are focused on the structure of the web considered in a large scale. The relationships among web domains have been analyzed in the Nordic academic web space (Ortega & Aguillo, 2008), or even in the world web space (Ortega & Aguillo, 2009), from the perspective of SNA. In Baeza-Yates and Castillo (2007) national web domains are analyzed attending to several criteria, in particular, degree and ranking. Page reputation is another topic related to link analysis frequently reported in the literature. In this case, SNA has also been applied considering the Indegree method as an alternative to Pagerank methods (Berlt et al., 2010). Finally, link analysis through SNA has been combined with text analysis to improve web information retrieval algorithms (Almpanidis, Kotropoulo, & Pitas, 2007).

Although Web structure has frequently been studied, comparatively little is known at the website level concerning its structure as an information organization and access mechanism. In this paper, an exploratory study for the identification of website link structure using factor analysis is proposed. For this purpose, the hypertext structures of eighty institutional websites have been extracted both at a domain and at a page level. Therefore, websites 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 considering 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. However, due to the exploratory nature of this study, it is difficult to select a subset of indicators to

* Corresponding author. Tel.: +34 954 48 12 93; fax: +34 954 48 73 73.
E-mail addresses: rmtorres@us.es (M.R. Martínez-Torres), toral@esi.us.es (S.L. Toral), beatriz@us.es (B. Palacios), fbarrero@esi.us.es (F. Barrero).
[1] Tel.: +34 954 55 43 10; fax: +34 954 55 69 89.

perform factor analysis, and the alternative of considering all possible subset of indicators is computationally prohibitive. As a solution, a genetic search of an optimum subset of indicators using a multi-objective fit function is proposed. The obtained result provides new insights about web site patterns and highlights the utility of genetic algorithms as a tool for new knowledge discovery.

The rest of the paper is structured as follows: a brief description of the methodology is provided in Section 2. In particular, network modeling of website structure, SNA features of extracted networks and factor analysis methodology are described. Section 3 is devoted to the application of genetic algorithms to the problem of extracting an optimum subset of variables able to explain the latent dimensions of website structure. The case study and results are discussed in Section 4. Finally, conclusions are detailed in Section 5.

2. Website structure analysis using social network analysis

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 links to other domains or pages outside the root domain are considered, the out-links from them will not be followed.

2.1. Social network analysis

A social network can be represented as a graph $G = (V,E)$ where V denotes a finite set of vertices and E denotes a finite set of edges such that $E \subseteq V \times V$. Some network analysis methods are easier to understand when graphs are conceptualized as matrices (Martínez-Torres, Toral, & Barrero, 2010; Nooy, Mrvar, & Batagelj, 2005) as shown in Eq. (1).

$$M = (m_{i,j})_{n*n} \quad \text{where } n = |V|, \quad m_{i,j} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In case of a valued graph, real valued weight function w(e) is defined on the set of edges, i.e. $w(e) = Ex\mathscr{R}$, and the matrix is then defined as given by Eq. (2).

$$m_{i,j} = \begin{cases} w(e) & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In the context of link analysis, the referred domain network is a star-shaped 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 lines. 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 refers to the number of lines and degree refers to the number of ties in which each vertex is involved.

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

- Size: the number of nodes represents the number of web pages and the 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 website structure and is higher than the depth of five used in some previous studies (Yang & Qin, 2008).
- Density: it is defined as the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines. The main problem of this definition is that it does not take into account valued lines 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 lines incident with it (Toral, Martínez-Torres, & Barrero, 2009). A higher degree of nodes yields a denser network, because nodes entertain more ties, and the average degree is a non-size dependent measure of density. 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 vertices 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. 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: it is defined as the number of steps in the shortest path that connects 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: it 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 vertices divided by the maximum variation in closeness centrality scores possible in a network of the same size (Toral, Martínez-Torres, Barrero, & Cortés, 2009).
- Betweenness: it 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 vertex is the proportion of all geodesics between pairs of other vertices that include this vertex, and betweenness centralization of the network is the variation in the betweenness centrality of vertices 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 to detect gateways connecting separate sub networks (Faba-Pérez, Zapico-Alonso, Guerrero-Bote, & Moya-Anegón, 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*-neighbour partition, in which nodes are clustered using the distance to the root node. The second one if 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* measures the statistical dependence between two classifications. Rajski's indices measure the degree to which the information in one classification is preserved in the other classification. Only the symmetrical version of Rajski's indices has been considered.

2.2. Factor analysis

Factor Analysis is a way to fit a model to multivariate data, estimating their interdependence. It addresses the problem of analyzing the structure of interrelationships among a number of variables by defining a set of common underlying dimensions, the factors, which are not directly observable, segmenting a sample into relatively homogeneous segments (Rencher, 2002). Because each factor may affect several variables in common, they are known as "common factors". Each variable is assumed to be dependent on a linear combination of the common factors, and the coefficients are known as loadings (Toral & Martínez Torres, 2009). Mathematically, the factor analysis model expresses each descriptor as a linear combination of underlying common factors $f_1, f_2, \ldots, f_m$, with an accompanying error term to account for that part of the variable that is unique (not in common with the other variables). For $y_1, y_2, \ldots, y_p$ in any observation vector y, the model is as follows:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m + \varepsilon_1 \\ y_2 - \mu_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots + \lambda_{2m}f_m + \varepsilon_2 \\ &\cdots \\ y_p - \mu_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \cdots + \lambda_{pm}f_m + \varepsilon_p \end{aligned} \tag{3}$$

Model (3) can be written in matrix notation as in Eq. (4), where Λ is the factor loadings matrix.

$$y - \mu = \Lambda f + \varepsilon \tag{4}$$

Ideally, m should be substantially smaller than p; otherwise we have not achieved a parsimonious description of the variables as functions of a few underlying factors. The coefficients λ_{ij} are called loadings and serve as weights, showing how each y_i individually depends on the underlying factors (Lee & Lee, 2011).

With appropriate assumptions, λ_{ij} indicates the importance of the jth factor f_j to the ith variable y_i and can be used in interpretation of f_j. For instance, f_2 could be interpreted by examining its coefficients, $\lambda_{12}, \lambda_{22}, \ldots, \lambda_{p2}$.

The larger loadings relate f_2 to the corresponding y's. From these y's, a meaning or description of f_2 could be inferred. It is expected the loadings will partition the variables into groups corresponding to factors.

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. The exploratory nature of this study has several implications:

- A high number of indicators related to SNA have been extracted for the two networks considered. The reduced theoretical background does not allow screening out unimportant indicators before analysis factor begins.
- The number of latent factors is unknown. Again, the lack of sufficient theoretical background means factors should be selected attending to the homogeneity of their indicators.

Next section proposes the use of genetic algorithms for searching an optimum solution and solving these problems. 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. Fortunately, an interesting property of loadings is that they can be multiplied by an orthogonal matrix preserving the essential properties of the original loadings. Let T be an arbitrary orthogonal matrix, TT' = I. Inserting TT' into the basic model (4):

$$y - \mu = \Lambda TT'f + \varepsilon \tag{5}$$

Associating T with Λ and T' with f, the model becomes:

$$y - \mu = \Lambda^* f^* + \varepsilon \text{ with } \Lambda^* = \Lambda T \text{ and } f^* = T^* f \tag{6}$$

It can be demonstrated that the new loadings $\Lambda^* = \Lambda T$ reproduce the covariance matrix (Rencher, 2002). This property is frequently used to facilitate the interpretation of factors. If we can achieve a rotation in which every point is close to an axis, then each variable loads highly on the factor corresponding to the axis and has small loadings on the remaining factors. In this case, there is no ambiguity. The rotated factor analysis fit ensures that factors represent unidimensional constructs.

3. Genetic search of website latent dimensions

A Genetic algorithm (GA) is a computational abstraction of biological evolution which can be used to solve some optimization problems. The technique was first introduced by Holland (1975) for use in adaptative systems. It is an iterative process which applies a series of genetic operators such as selection, crossover and mutation to a population of elements. These elements, called chromosomes or individuals, represent possible solutions to the problem. The initial population is randomly selected from the solution space. Genetic operators combine the genetic information of the elements to form new generations of populations. Each chromosome has an associated fitness value which quantifies its value as a solution to the problem. The chromosomes compete to reproduce based on their fitness values, thus the chromosomes representing better solutions have a higher chance of survival. The crossover involves two chromosomes whose portions are swapped. Selection according to fitness combined with crossover gives the GA its evolutionary power. The GA uses an elitist strategy meaning that the best individual is carried over to the next generation so that we can only improve the solution over the course of the genetic optimization. The algorithm stops when some stopping criterions are satisfied (Martínez-Torres & Toral, 2010). Several questions should be taking into account when applying GA:

- Chromosomal encoding, how to represent possible solutions.
- Fitness function selection. It must accurately represent the value of the solution.
- Parameter values selection (population size, number of iterations, probabilities, etc.)

In this study, the use of GA is justified due to its exploratory nature. A total to 64 indicators (see appendix) have been extracted according to the SNA features detailed in Section 2.1. The problem of choosing a subset of indicators leading to interpretable latent factors is unaffordable when trying to explore all the possibilities. Notice that the space of possible solutions is formed by 2^{64} = 1.8447e+019 possibilities. That means that we should perform 2^{64} different factor analyses to completely explore the space of possible solutions. In this kind of problems, GA can perform a guided search of the optimum solution with lower computational cost than exploring one by one all the possibilities.

The first condition to apply GA properly is a good selection of the chromosomal encoding, which should be valid and complete. Our chromosomal encoding is constituted by a 64 binary sequence in which "ones" are the variables that are going to be used in factor analysis, and "zeros" represents variables that are going to be excluded from this analysis. Clearly, the encoding representation is complete, as the 2^{64} possibilities are able to be represented, and valid, as all of them can be computed.

The next step is the fitness function selection. The fitness function quantifies the suitability of each chromosome as a solution. Chromosomes with high fitness have more chance of being selected, passing their genetic material (via reproduction or crossover) to the next generation. The fitness function provides the pressure for evolution towards a new generation with chromosomes of higher fitness than the previous ones. The chromosome representing the optimal solution should have the maximum fitness value for the solution space, and near optimal solutions should have higher fitness values. In the context of factor analysis, it is not possible to build a simple fitness function (Liu, Chen, & Chou, 2010). Fitness function should be multi-objective fitness function considering several parameters, like explained variance, correlations and interpretability of the latent factors.

$$F = c_1 Var + c_2 \frac{1}{n} \sum_{i=1}^{k} r_i^2 + c_3 Interp \tag{7}$$

- Explained variance (*Var*). Factor analysis results show the explained variance by the considered factors (usually, the number of factors is given by the number of eigenvalues of the correlation data matrix bigger than 1). The explained variance through the selected number of indicators should be maximized. But it is not the unique parameter to be taken into account. A fitness function equal to the explained variance will tend to the trivial solution of just considering one indicator. This is due to the fact that it is easier to explain the variance of a data set when it is formed by a small number of data.
- Correlations between variables ($1/n \sum_{i=1}^{k} r_i^2$). The average of the sum of the squared correlation coefficients between indicators is used as the second part of the fitness function. This term will tend by itself to the trivial solution of considering the whole data set. It is the reverse strength to the previous part of the fitness function.
- Interpretability of factors. The third part of the fitness function penalizes factors with less than three indicators. The reason for choosing the value of 3 is because factors explained with less than three indicators are not considered well-defined in the literature (Rencher, 2002). This part of the fitness function is the most important one as it is promoting a reduced number of factors with more indicators, improving the final interpretation of the latent factors.

$C1$, $C2$, and $C3$ coefficients are used to adjust the relative importance of the three parts of the fitness function. Obviously, the range of them is [0, 1], with the restriction of $C1 + C2 + C3 = 1$.

The final decision for GA application refers to parameter values selection. GA performance may be sensitive to certain parameter values, particularly the population size, the frequency of operator selection and the termination criterion. All of them vary considerably, and there is little or no documented justification for their selection. Nevertheless, a high value for the population size may reduce this sensibility to GA parameters. In this paper, population size has been chosen equal to 10000, with a 20% of reproduction rate. The value of 10000 is considered a good value to obtain richness of genetic content. These values are typical in the literature about GA (Goldberg, 1989; Martínez-Torres & Toral, 2010).

4. Case study

The genetic search of web sites latent dimensions has been applied to 80 Spanish University web sites. All of them are included in the Webometrics Ranking of World Universities (www.webometrics.org), where more than 6000 universities all over the world are sorted according to size and visibility. Table 1 lists the root domains of the considered web sites. They cover almost the whole range of Webometrics Ranking, and exhibit a variety of size in term of domains and web pages. Table 2 summarizes some descriptive statistics. The first column shows that more than 718.000 web pages and more than four million outlinks have been considered. Figs. 1 and 2 shows the particular case of the domain and page network, respectively, corresponding to the particular case of the University of Seville. For each web site, two starting networks have been collected: the domain network and the page network.

The social network features of Section 2.1 have been measured, considering in some cases the whole network, and in some cases

Table 1
List of considered web sites.

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

Table 2
Websites Descriptive statistics.

	Sum	Mean	SD
Subdomains	2438	30,47	38,10
Ext. domains	30500	381,25	580,32
Pages	718272	8978,40	15334,01
Out-links	4429231	55365,38	73290,17

the subnetworks excluding nodes with 0 out-degree or subnetworks with $k > 1$ cores. As a result, 64 indicators have been obtained.

4.1. Data analysis

GA has been applied to obtain an optimum subset of indicators able to identify web site profiles according to their link structure. The cost function follows the general structure defined in Section

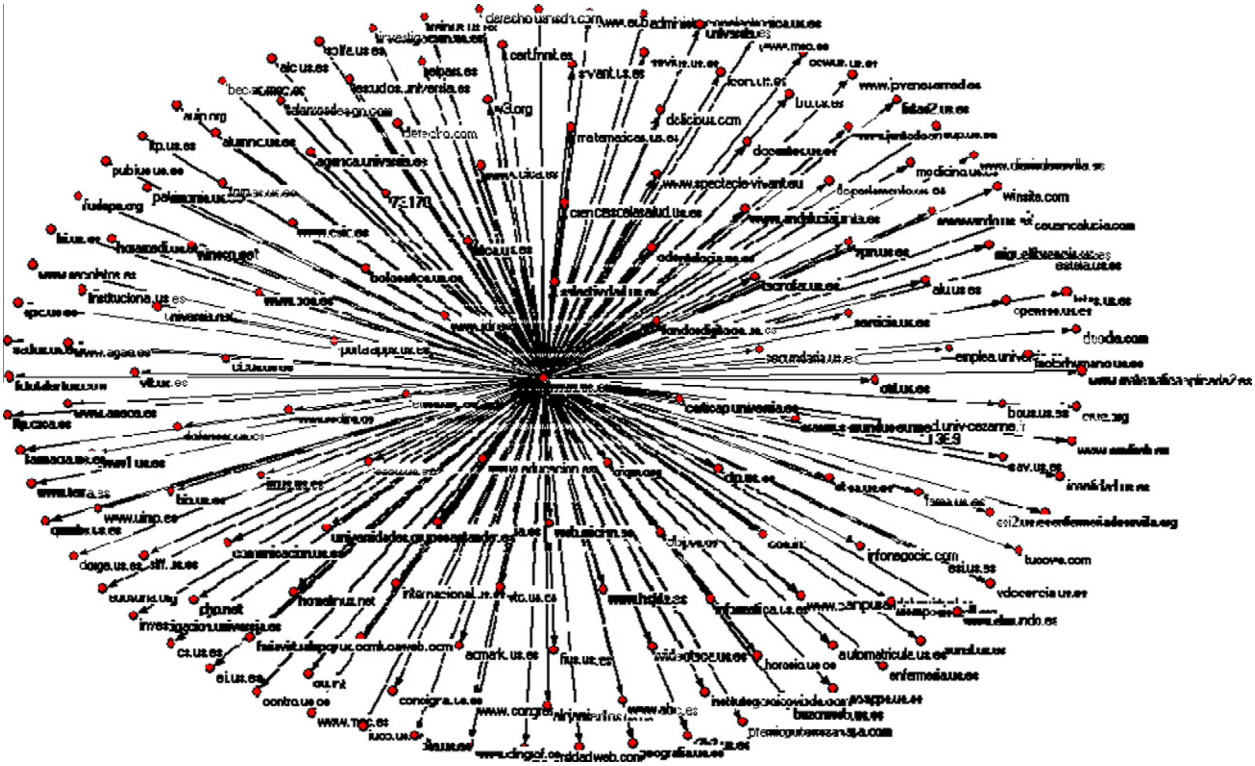


Fig. 1. University of Seville domain network.

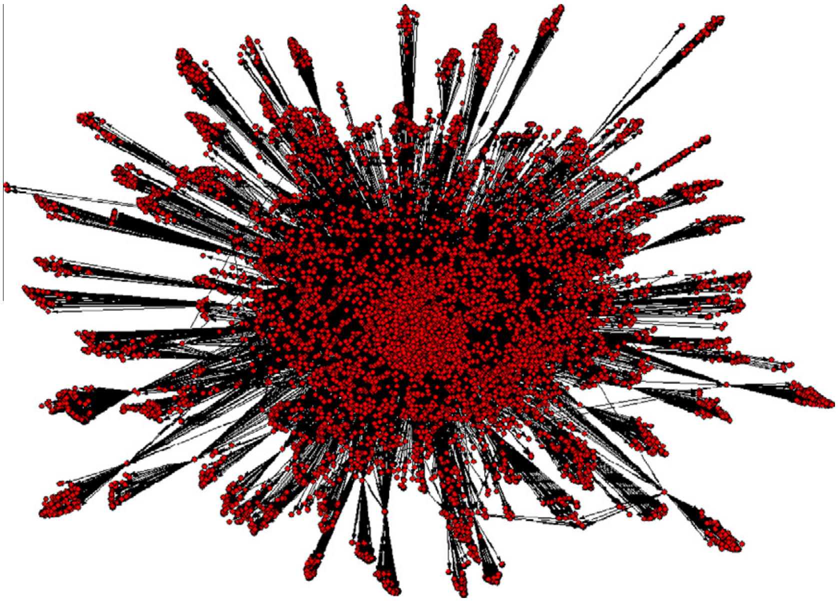
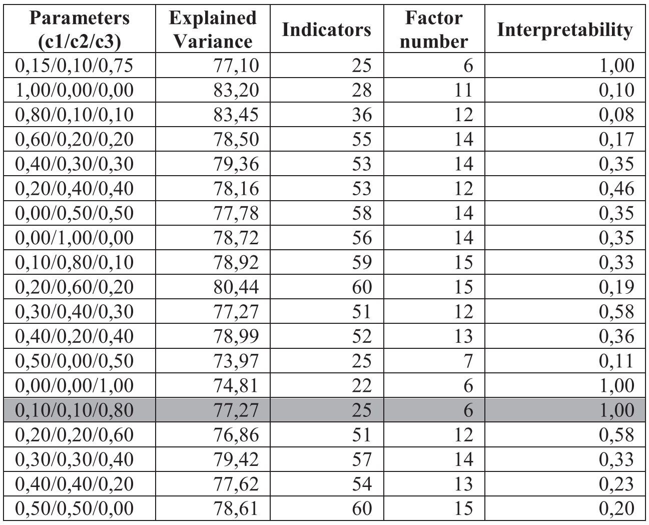


Fig. 2. University of Seville page network.

Table 3
Sensitivity analysis for different $c1$, $c2$ and $c3$ values.



Parameters (c1/c2/c3)	Explained Variance	Indicators	Factor number	Interpretability
0,15/0,10/0,75	77,10	25	6	1,00
1,00/0,00/0,00	83,20	28	11	0,10
0,80/0,10/0,10	83,45	36	12	0,08
0,60/0,20/0,20	78,50	55	14	0,17
0,40/0,30/0,30	79,36	53	14	0,35
0,20/0,40/0,40	78,16	53	12	0,46
0,00/0,50/0,50	77,78	58	14	0,35
0,00/1,00/0,00	78,72	56	14	0,35
0,10/0,80/0,10	78,92	59	15	0,33
0,20/0,60/0,20	80,44	60	15	0,19
0,30/0,40/0,30	77,27	51	12	0,58
0,40/0,20/0,40	78,99	52	13	0,36
0,50/0,00/0,50	73,97	25	7	0,11
0,00/0,00/1,00	74,81	22	6	1,00
0,10/0,10/0,80	77,27	25	6	1,00
0,20/0,20/0,60	76,86	51	12	0,58
0,30/0,30/0,40	79,42	57	14	0,33
0,40/0,40/0,20	77,62	54	13	0,23
0,50/0,50/0,00	78,61	60	15	0,20

Table 4
Selected subset of indicators.

	Indicator	Network
I1	External domains	Domain Net.
I2	Average degree	Domain Net.
I3	Density	Domain Net.
I4	Number of pages	Page Net.
I5	Number of pages in the last level (depth of 7)	Page Net.
I6	Number of no-returning pages (excluding last level)	Page Net.
I7	Out-degree standard deviation	Page Net.
I8	Number of strong components	Page Net.
I9	% of pages included in strong components	Page Net.
I10	*K*-core including the maximum number of pages	Page Net.
I11	Average value of closeness centrality	Page Net.
I12	Standard deviation of closeness centrality	Page Net.
I13	Number of pages	Page Net. excluding out-degree = 0
I14	Betweeness centralization	Page Net.
I15	Standard deviation of egocentric density	Page Net.
I16	Average value of nodes betweeness centrality	Page Network of *k*-cores, *k* > 0
I17	Standard deviation of vertices betweeness centrality	Page Network of *k*-cores, *k* > 0
I18	Average value of egocentric density	Page Network of *k*-cores, *k* > 0
I19	Average value of vertices betweeness centrality	Page Net. excluding out-degree = 0
I20	Average value of egocentric density	Page Net. excluding out-degree = 0
I21	Number of vertices developing a brokerage role	Page Net. excluding out-degree = 0
I22	Standard deviation of brokerage roles	Page Net. excluding out-degree = 0
I23	Cramer's V index of partition correlation (out-degree, *k*-neighbour)	Page Net.
I24	Rajski's index of partition correlation (out-degree, *k*-neighbour)	Page Net.
I25	Rajski's index of partition correlation (out-degree, *k*-neighbour)	Page Net. excluding out-degree = 0

3 but considering the values $c1 = 0, 1$, $c2 = 0, 1$ and $c3 = 0, 8$. Notice that interpretability of factors has been clearly overweighed. This strategy seems reasonable, since factors with less than three indicators are not admissible in factor analysis. Besides, interpretability guides GA towards a reduced number of factors, which is also reasonable to find factors with clear and separate meanings. Table 3 details obtained results when running GA for different parameters values. This table shows that interpretability only reaches a value of 1 when $c3$ is overweighed, while the explained variance remains above 70% in all the cases. It is important to achieve at least three indicators per factor. Otherwise, factors are not well defined. The results for the selected values of $c1$, $c2$ and $c3$ are highlighted in dark gray in Table 3.

Using the optimum parameters values, GA has converged after 30 generations, with an explained variance of 77,10%, and 25 indicators grouped in 6 factors. All of them include at least three indicators, and their meaning, using Varimax rotation, are interpretable. Time required by genetic algorithm execution is 4.822,49 s (80,37 min). This value is much smaller than the alternative option of exploring the whole solution space. Taking into account that each factor analysis requires 12.9 ms, the $2^{64} = 1.8447e+019$ possibilities of the solution space would require millions of years. The selected subset of indicators is listed in Table 4. In particular, the indicators description and the network over which it is calculated are detailed.

The evolution of the genetic clustering algorithm is detailed in Fig. 3. The initial population (generation 0) has a low fitness value, which indicates that the individuals of the population are far from the optimum. As the number of generations increase, the fitness of individuals within the population also increases, as the genetic algorithm is biased towards the survival of genetic material contained within the individuals with high fitness function values.

The results from factor analysis using the set of variables selected by the genetic algorithm are detailed in Table 5. Usually, a

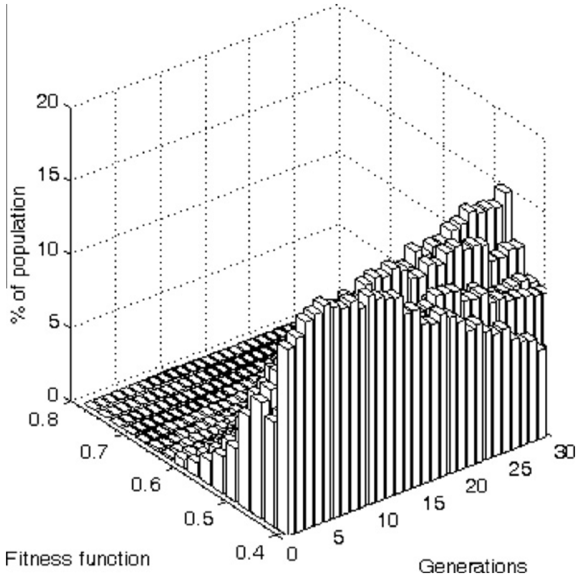


Fig. 3. Fitness distribution over 30 generations of the genetic algorithm.

Table 5
Explained variance of resulting factor analysis.

Factor	Eigenvalues		
	Value	% Variance	% Cumulative
1	7,990	31,962	31,962
2	3,852	15,407	47,369
3	2,911	11,646	59,015
4	1,857	7,427	66,442
5	1,656	6,624	73,065
6	1,010	4,039	77,104
7	,833	3,333	80,437
.	.	.	.
.	.	.	.
.	.	.	.
25	,007	,029	100,000

number of factors equal to the number of eigenvalues higher than 1 is selected (Rencher, 2002). Consequently, up to 6 latent factors can be distinguished as result of factor analysis.

Table 7
Statistical significance of ANOVA.

	F	Sig		F	Sig
I1	7,376	,000	I14	11,221	,000
I2	3,823	,002	I15	7,816	,000
I3	3,754	,003	I16	21,161	,000
I4	19,282	,000	I17	21,805	,000
I5	17,323	,000	I18	6,664	,000
I6	8,250	,000	I19	13,452	,000
I7	7,505	,000	I20	7,888	,000
I8	28,878	,000	I21	8,612	,000
I9	6,171	,000	I22	1,219	,306
I10	5,912	,000	I23	8,048	,000
I11	7,059	,000	I24	4,626	,000
I12	5,941	,000	I25	5,692	,000
I13	15,354	,000			

The indicators associated to each factor are obtained from the factor loadings using a Varimax rotation. All the indicators associated in this way with the same factor are hypothesized to share a common meaning that the analyst should discover. Table 6 shows which indicators are associated to each factor and their corresponding factor loading.

On the other hand, factor scores are used to categorize the original sample of Universities, which can be approximated to one of the identified latent factors. Consequently, the original sample of Spanish University websites can be categorized in six groups. An analysis of variance (ANOVA) has been applied to the categorization of the original sample in the six groups obtained form factor analysis. The aim of this analysis consists of checking the null hypothesis of equal population means. Table 7 details the F statistic, the ratio of two different estimators of population variance, which appears together with its corresponding critical level or observed significance. The results is that the null hypotheses have been rejected in all the cases with a significance value below 0,05. That means the obtained categorization from factor analysis is well defined.

Table 8 details the mean value of the considered 25 indicators per each of the distinguished groups. Using this information as well as the factor loadings of Table 6, the following websites structure patterns can be distinguished:

Table 6
Identified factors.

		Description	Loading
F1	I2	Average degree	−0,724
	I16	Average value of nodes betweeness centrality	0,903
	I17	Standard deviation of vertices betweeness centrality	0,884
	I19	Average value of vertices betweeness centrality	0,839
	I23	Cramer's V index of partition correlation (out-degree, *k*-neighbour)	0,703
	I25	Rajski's index of partition correlation (out-degree, *k*-neighbour)	0,746
F2	I9	% of pages included in strong components	0,722
	I11	Average value of closeness centrality	0,924
	I12	Standard deviation of closeness centrality	0,718
	I14	Betweeness centralization	0,826
	I24	Rajski's index of partition correlation (out-degree, *k*-neighbour)	0,578
F3	I15	Standard deviation of egocentric density	0,763
	I18	Average value of egocentric density	0,895
	I20	Average value of egocentric density	0,875
F4	I4	Number of pages	0,900
	I5	Number of pages in the last level (depth of 7)	0,928
	I13	Number of pages	0,661
	I21	Number of vertices developing a brokerage role	0,510
F5	I1	External domains	0,852
	I6	Number of no-returning pages (excluding last level)	0,647
	I8	Number of strong components	0,831
F6	I7	Out-degree standard deviation	0,786
	I10	*K*-core including the maximum number of pages	0,633
	I22	Standard deviation of brokerage roles	0,635

Table 8
Mean values of selected indicators.

	F1	F2	F3	F4	F5	F6
I1	135,733	227,000	393,263	303,667	1479,000	362,600
I2	1,952	1,972	1,993	1,992	1,998	1,993
I3	0,119	0,139	0,036	0,040	0,015	0,024
I4	1871,600	2929,308	7436,947	65815,667	14504,857	11386,600
I5	242,800	138,846	299,895	50867,000	325,286	1549,667
I6	1415,267	1835,846	4769,790	9678,667	11348,286	7815,400
I7	14,214	17,557	17,673	17,509	12,663	28,117
I8	2,133	1,462	3,053	1,333	14,286	1,400
I9	6,487	28,075	6,697	5,241	0,161	11,580
I10	10,133	11,846	22,790	2,000	6,714	35,000
I11	0,054	0,115	0,049	0,014	0,035	0,042
I12	0,105	0,132	0,092	0,048	0,075	0,093
I13	213,533	954,615	1204,105	5270,000	2831,286	2022,200
I14	0,071	0,183	0,064	0,046	0,088	0,042
I15	0,237	0,301	0,298	0,115	0,226	0,244
I16	0,013	0,004	0,003	0,001	0,002	0,002
I17	0,048	0,024	0,018	0,015	0,013	0,012
I18	0,569	0,580	0,679	0,352	0,513	0,581
I19	0,010	0,003	0,003	0,001	0,002	0,002
I20	0,524	0,570	0,661	0,308	0,473	0,548
I21	98,533	559,615	544,632	2614,000	1333,286	1438,467
I22	269,061	2431,400	1687,299	2740,362	5774,782	11390,647
I23	0,506	0,455	0,358	0,344	0,291	0,334
I24	0,137	0,135	0,079	0,180	0,052	0,071
I25	0,262	0,159	0,127	0,098	0,067	0,100

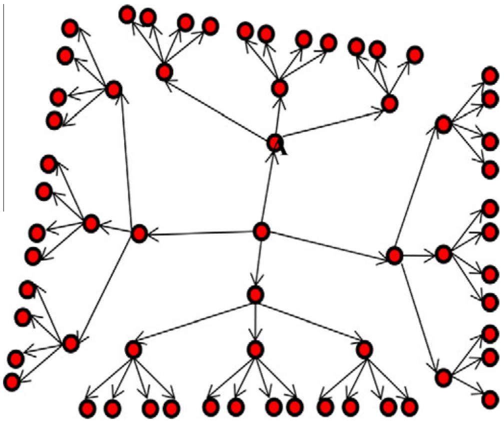


Fig. 6. Symbolic representation of factor 3 websites.

Factor 1 represents a distributed structure of the website, with a lot of nodes developing a betweenness role. The high value of partition correlations also supports the distributed structure with lower and intermediate level pages (near the root domain) acting as directories of information and higher level pages (far from the root domain) providing more detailed information. The high value of Rajski (I25) and Cramer's V (I23) information indices indicates the out-degree is growing as vertices are more distant from the root domain. The high value of average value and standard deviation of vertices betweeness centrality (I16, I17 and I19) suggest the website is structured through highly interconnected vertices spread over the website, following a certain tree structure. Fig. 4 is a symbolic representation of factor 1 websites.

Factor 2 represents a more centralized structure in the sense of distance to the root domain. There is a core of highly interconnected pages, but the information is also spread out as we move toward deeper levels in the structure. The length of the path required to access any node of the network is the shortest among all the distinguished website patterns. This characteristic is supported by the high value of average and standard deviation of closeness centrality, which suggest a flat-shaped structure. The symbolic representation of Fig. 5 illustrates this kind of structure. The distance to reach node B from the root domain A is two, as a difference to factor 1 symbolic representation, where a longer distance of 4 is shown.

Factor 3 refers to an egocentric structure, where the global network could be considered as the sum of more or less independent subnetworks, that is, one may look at the structure of local networks within a complete network. This factor represents websites with a clear division in independent areas, with a low inter-connection among them (Fig. 6).

Factor 4 considers large web sites. The number of pages grows geometrically with the depth level, so it is necessary a long navigation process to achieve the desired information (see Fig. 7). Just the opposite to factor 1, the low value of Rajski (I25) and Cramer's V (I23) information indices indicates a low structured website. Consequently, this factor describes a full network structure where a page is linked to a lot of other pages in the Web site; this enables visitors to navigate through the available information as they wish, but at the cost of complexity. Although the hyperlink navigation

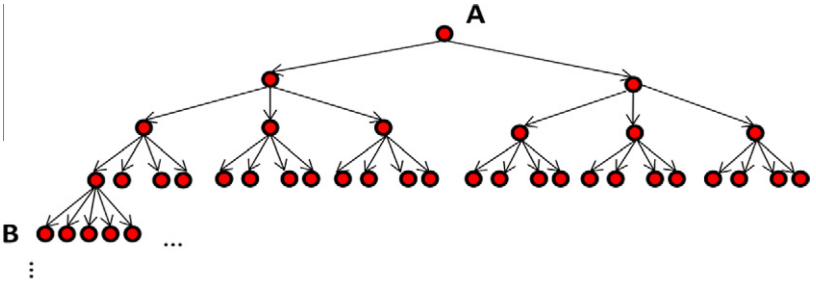


Fig. 4. Symbolic representation of factor 1 websites.

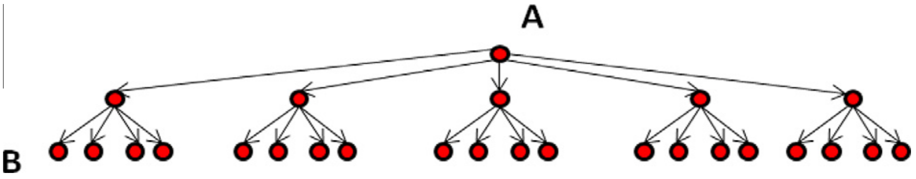


Fig. 5. Symbolic representation of factor 2 websites.

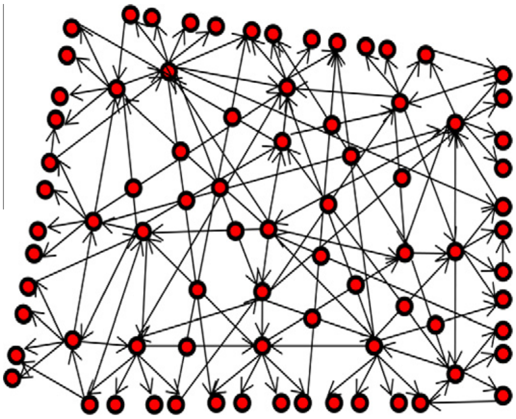


Fig. 7. Symbolic representation of factor 4 websites.

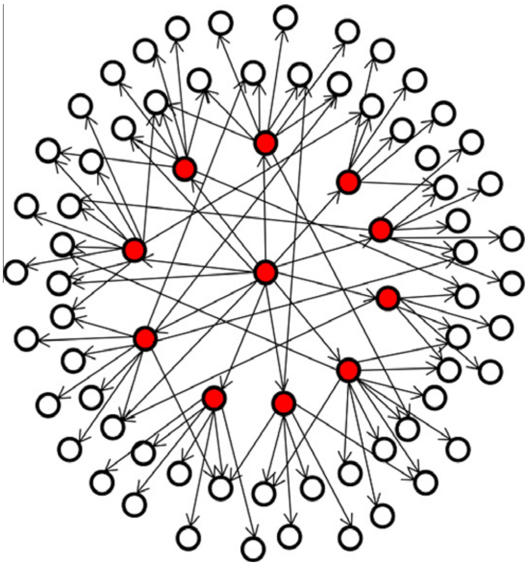


Fig. 8. Symbolic representation of factor 5 websites.

structure might become cumbersome, this problem is alleviated with search functions leading users directly to the requested information.

Factor 5 represents smaller web sites, where a great amount of information is provided using external references to the website. This idea is supported by the high average value of non-returning pages excluding pages located in the last level.

Fig. 8 is a symbolic representation of websites identified by factor 5, with dark nodes representing the internal domains and white nodes representing external domains.

Factor 6 represents web site with a structure dominated by one subnetwork, containing the most relevant information. This pattern is supported by the high value of Indicator I10, related to *k*-cores. A *k*-core indentifies a kernel of degree *k*, that is, a maximum subgraph in which all points are adjacent to at least *k* other points. This kernel constitutes the dominant subnetwork of the website (see Fig. 9).

Basically, the identified profiles of web site structures respond to two basic strategies when deciding their final structure (Tan & Wei, 2006). The first strategy consists of offering a structure which makes sense to the final user. In this sense, web sites sacrifices accessibility of information looking for a more structured navigation scheme. Factors 1 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 Website is kept to a moderate level (Tan & Wei, 2006). In this sense, factor 2 represents ordered but flat-structured websites for improving accessibility. Finally, factor 4, 5 and 6 could be considered as a mixture of both strategies. They exhibit a lack of a defined structure but a lot of paths among nodes guarantee the required information can be easily accessed.

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 network (Huizingh, 2000). Several of their proposed navigation structure types have been split in this study. For instance, the tree structure leads to a depth tree structure (factor 1), a flat-tree structure (factor 2) and a networked tree structure (factor 3), and extensive network type lead to large websites (factor 4) and websites dominated by a *k*-core structure (factor 6). External domains were not considered in this previous study.

Website internal structure is strongly related to issues like accessibility and navigability through websites. Navigation features allow the site visitor easy access to information of interest, both internal and external to the site, and it is included as one of the design features of corporate websites, along with presentation, security, speed and tracking (Robbins & Stylianou, 2003). The

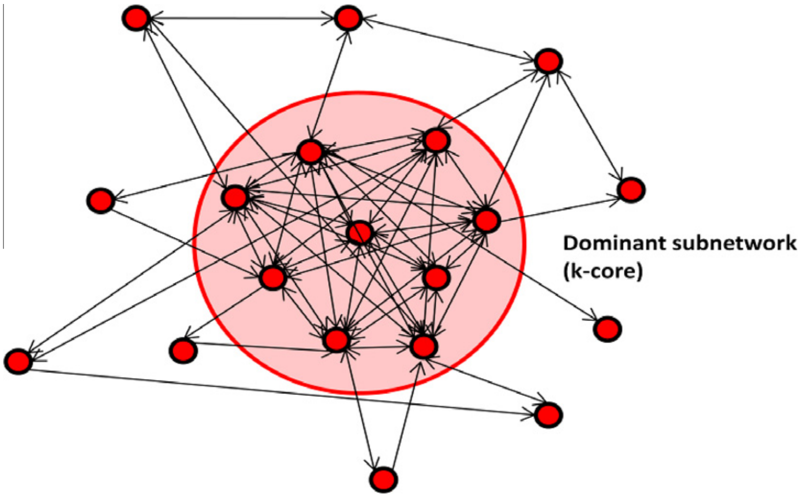


Fig. 9. Symbolic representation of factor 6 websites.

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 & Bañegil, 2004).

5. Conclusion

This paper has developed a tool for identifying website link structures considering websites as social networks. The use of evolutionary computation techniques has allowed extracting the main profiles in the particular case of institutional websites from Spanish Universities. Obtained results agree with the general rules of website designs proposed in the literature, and they are useful for web designers and organizations when taking decisions about their web presence and their corporate image. Although the study is limited to Spanish Universities Websites, they constitute a rich enough sample among the Webometrics Ranking of World Universities. This study could be extended to other institutional web sites to validate the obtained results.

Acknowledgment

This work has been supported by the Spanish Ministry of Education and Science (Research Project with reference DPI2007-60128) and the Consejería de Innovación, Ciencia y Empresa (Research Project with reference P07-TIC-02621).

Appendix A

Selected indicators	Indicators	Description	Network
	VAR01	Number of subdomains	Domain network
I1	VAR02	Number of external domains	
	VAR03	Density	
I2	VAR04	Average degree	
	VAR05	Number of lines with value = 1	Domain network (excluding external domains)
I3	VAR06	Density	
	VAR07	Average degree	
I4	VAR08	Number of pages	Page network (excluding external pages)
	VAR09	Total number of lines	
	VAR10	Number of lines with value > 1	
	VAR11	Density	
I5	VAR12	Number of pages in the last level	
I6	VAR13	Number of non returning pages (excluding pages of the last level)	
	VAR14	Average out-degree	
I7	VAR15	Standard deviation out-degree	
I8	VAR16	Number of strong components	
I9	VAR17	% of pages included in strong components	
	VAR18	maximum k-core	
	VAR19	Number of pages included in the maximum k-core	
I10	VAR20	K-core including the maximum number of pages	
	VAR21	Number of cores with $k > 0$	
	VAR22	% of pages included in k-cores ($k > 0$)	
I11	VAR23	Average value of closeness centrality	
I12	VAR24	Standard deviation of closeness centrality	
I13	VAR25	Number of pages (excluyendo externas y degree output $1 - *$)	Page network (excluding external pages + degree output $1 - *$)
	VAR26	Total number of lines	
	VAR27	Number of lines with value > 1	
	VAR28	Density	
	VAR29	Average degree	
	VAR30	Average out-degree (media aritmetica)	
	VAR31	Standard deviation out-degree	
	VAR32	% of pages included in strong components	
	VAR33	Number of cores with $k > 0$	
	VAR34	% of pages included in k-cores ($k > 0$)	
	VAR35	Average value of closeness centrality	
	VAR36	Standard deviation of closeness centrality	
	VAR37	Average value of closeness centrality	Page network of k-cores, k>0
	VAR38	Standard deviation of closeness centrality	
I14	VAR39	Betweeness centralization	Page network (excluding external pages)
	VAR40	Average value of vertices betweeness centrality	
	VAR41	Number of vertices with betweeness centrality > 0	
	VAR42	egocentric density (average value)	
I15	VAR43	egocentric density (SD)	
	VAR44	Betweeness centralization	Page network of k-cores, $k > 0$
I16	VAR45	Average value of vertices betweeness	

Appendix A (*continued*)

Selected indicators	Indicators	Description	Network
		centrality	
I17	VAR46	Standard deviation of vertices betweeness centrality	
I18	VAR47	egocentric density (average value)	
	VAR48	egocentric density (SD)	
	VAR49	Betweeness centralization	Page network (excluding vertices with out-degree = 0)
I19	VA050	Average value of vertices betweeness centrality	
	VAR51	Standard deviation of vertices betweeness centrality	
I20	VAR52	egocentric density (average value)	
	VAR53	egocentric density (SD)	
I21	VAR54	Number of vertices developing a brokerage role	
	VAR55	Average brokerage role	
I22	VAR56	Standar deviation	
	VAR57	Spearman Rank Correlation Coefficient	Page network (excluding external pages)
	VAR58	Pearson Correlation Coefficient	
I23	VAR59	Cramer's V	
I24	VAR60	Rajski (C1 ↔ C2)	
	VAR61	Spearman Rank Correlation Coefficient	Page network (excluding vertices with out-degree = 0)
	VAR62	Pearson Correlation Coefficient	
	VAR63	Cramer's V	
I25	VAR64	Rajski (C1 ↔ C2)	

References

Almpanidis, G., Kotropoulo, C., & Pitas, I. (2007). Combining text and link analysis for focused crawling – An application for vertical search engines. *Information Systems, 32*, 886–908.

Baeza-Yates, R., & Castillo, C. (2007). Characterization of national web domains. *ACM Transactions on Internet Technology, 7*(2), 1–32.

Berlt, K., Silva de Moura, E., Carvalho, A., Cristo, M., Ziviani, N., & Couto, T. (2010). Modeling the web as a hypergraph to compute page reputation. *Information Systems, 35*(5), 530–543.

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

Faba-Pérez, C., Zapico-Alonso, F., Guerrero-Bote, V. P., & Moya-Anegón, F. (2005). Comparative analysis of webometric measurements in thematic environments. *Journal of the American Society for Information Science and Technology, 56*(8), 779–785.

Goldberg, D. A. (1989). *Genetic algorithm – in search, optimization and machine learning*. Addison-Wesley Publishing Company, Inc.

Holland, J. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Huizingh, E. K. (2000). The content and design of web sites: An empirical study. *Information & Management, 37*(3), 123–134.

Iacobucci, D. (1994). Graphs and matrices. In S. Wasserman & K. Faust (Eds.), *Social network analysis – methods and applications* (pp. 2–166). New York: Cambridge University Press.

Lee, Y., & Lee, H. (2011). Application of factor analysis for service R&D classification: A case study on the Korean ICT industry. *Expert Systems with Applications, 3*(3), 2119–2124.

Liu, T.-K., Chen, C.-H., & Chou, J.-H. (2010). Optimization of short-haul aircraft schedule recovery problems using a hybrid multiobjective genetic algorithm. *Expert Systems with Applications, 37*(3), 2307–2315.

Martínez-Torres, M. R., & Toral, S. L. (2010). Strategic group identification using evolutionary computation. *Expert Systems with Applications, 37*(7), 4948–4954.

Martínez-Torres, M. R., Toral, S. L., & Barrero, F. (2010). The role of internet in the development of future software projects. *Internet Research, 20*(1), 72–86.

Miranda González, F. J., & Bañegil, T. M. (2004). Quantitative evaluation of commercial web sites: An empirical study of Spanish firms. *International Journal of Information Management, 24*, 313–328.

Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory network analysis with pajek*. New York: Cambridge University Press.

Ortega, J. L., & Aguillo, I. F. (2008). Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing and Management, 44*(4), 1624–1633.

Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing and Management, 45*(2), 272–279.

Park, H. W., & Thelwall, M. (2003). Hyperlink analysis: Between networks and indicators. Journal of Computer-Mediated Communication 8,(4),http://www.ascusc.org/jcmc/vol8/issue4/park.html.

Rencher, A. C. (2002). *Methods of multivariate analysis. Wiley series in probability and statistics* (2nd ed.,). John Wiley & Sons.

Robbins, S. S., & Stylianou, A. C. (2003). Global corporate web sites: an empirical investigation of content and design. *Information & Management, 40*(3), 205–212.

Tan, G. W., & Wei, K. K. (2006). An empirical study of web browsing behaviour: Towards an effective website design. *Electronic Commerce Research and Applications, 5*(4), 261–271.

Thelwall, M. (2004). *Link analysis: An information science approach*. Amsterdam: Elsevier.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science, 34*(4), 605–621.

Toral, S. L., & Martínez Torres, M. R. (2009). International Comparison of R&D Investment By European, US and Japanese Companies. *International Journal of Technology Management, 49*(1/2/3), 107–122.

Toral, S. L., Martínez Torres, M. R., & Barrero, F. (2010). Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology, 52*(3), 296–303.

Toral, S. L., Martínez-Torres, M. R., & Barrero, F. (2009). Virtual Communities as a resource for the development of OSS projects: The case of Linux ports to embedded processors. *Behavior and Information Technology, 28*(5), 405–419.

Toral, S. L., Martínez-Torres, M. R., Barrero, F., & Cortés, F. (2009). An empirical study of the driving forces behind online communities. *Internet Research, 19*(4), 378–392.

Yang, B., & Qin, J. (2008). Data collection system for link analysis. In *Third International Conference on Digital Information Management, ICDIM*, (pp. 247–252).