2010 International Conference on Advances in Social Networks Analysis and Mining

Application of Genetic Algorithms to the Identification of Website Link Structure

M. R. Martínez Torres, B. Palacios, S. L. Toral, *Senior Member, IEEE*, and F. Barrero, *Senior Member, IEEE*

***Abstract*—This paper explores website link structure considering websites as interconnected graphs and analyzing their features as a social network. 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, a genetic search of their optimum number is proposed, and applied to a case study based on 80 Spanish University websites. Results provide coherent and relevant website profiles, and highlight the possibilities of Genetic Algorithms as a tool for discovering new knowledge related to website link structures.**

***Index Terms*—Link Analysis, Website structure, Factor Analysis, Genetic Algorithms.**

I. 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 [1], [2]. Web links have been heavily studied during the last years in order to understand the structure and growth patterns of the Web [3], 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 [3].

Social network analysis (SNA) has been frequently used for the study of link analysis [4], [5]. 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 [6]. 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 [7], or even in the world web space [8] from the perspective of SNA. In [9], 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 [10]. Finally, link analysis through SNA has been combined with text analysis to improve web information retrieval algorithms [11].

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 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 sites 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 II. In particular, network modeling of website structure, SNA features of extracted networks and factor analysis methodology are described. Section III is devoted to the application of genetic algorithms to the problem of extracting

Manuscript received December 9, 2009. 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)..

M. R. Martínez Torres and B. Palacios are with the Department of Business Administration and Marketing, University of Seville, Spain).

S. L. Toral, and F. Barrero are with the Electronic Engineering Department, University of Seville, Spain (phone: +34 954481293; fax: +34 954487373; e-mail: toral@esi.us.es)..

978-0-7695-4138-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ASONAM.2010.12

an optimum subset of variables able to explain the latent dimensions of website structure. The case study and results are discussed in section IV. Finally, conclusions are detailed in section V.

II. Website Structure Analysis using SNA

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.

A. SNA

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 ⊆ V × V. Some network analysis methods are easier to understand when graphs are conceptualized as matrices [12], as shown in Equation (1).

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

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

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

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 de 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 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 [13].
- 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 [14]. 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 [12].
- 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 [12]. 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 in [7] to detect sub-networks among Nordic academic web sites.
- Distance: it 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: 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 [15].
- 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 [12]. 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 [16].

- 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 [5]. Two important partitions can be extracted using network features previously introduced. The first one if 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 [12]. 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.

B. 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 [17]. 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 [18].

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:

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

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

$$y-\mu=\Lambda f+\varepsilon \quad (4)$$

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.

With appropriate assumptions, λ_{ij} indicates the importance of the j^{th} factor f_j to the i^{th} 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):

$$y-\mu=\Lambda TT'f+\varepsilon \quad (5)$$

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

$$y-\mu=\Lambda^* f^*+\varepsilon \text{ with } \Lambda^*=\Lambda T \text{ and } f^*=T'f \quad (6)$$

It can be demonstrated that the new loadings $\Lambda^*=\Lambda T$ reproduce the covariance matrix [17]. 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.

III. 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 [19] for use in adaptative systems. It is an iterative process which applies a series of genetic operators such us 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 [20]. 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. Up to 64 indicators has been extracted according to the SNA features detailed in section II.A. The problem of choosing a subset of indicators leading to interpretable latent factors is unaffordable when trying to explore all the possibilities. 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. Fitness function should be multi-objective fitness function considering several parameters, like explained variance, correlations and interpretability of the latent factors.

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

- 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 [17]. 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 genetic content. These values are typical in the literature about GA [20], [21].

IV. 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 I 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 II summarizes some descriptive statistics. The second column shows that more than 718.000 web pages and more than four million outlinks have been considered. Figure 1and Figure 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.

Table I. 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 II. 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 social network features of section II.A have been measured, considering in some cases the whole network, and in some cases the subnetworks excluding nodes with 0 out-degree or subnetworks with k>1 cores. As a result, 64 indicators have been obtained.

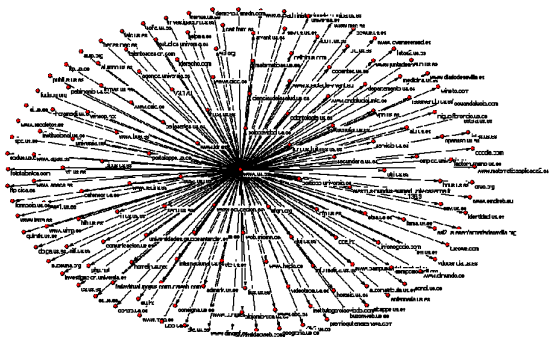

Figure 1. University of Seville domain network.

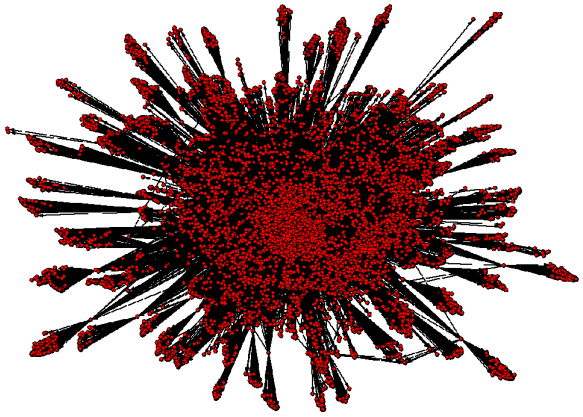

Figure 2. University of Seville page network.

A. 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 follow the general structure defined in section III but considering the values c1=0,15, c2=0,1 and c3= 0,75. 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.

Beginning with an initial randomly generated population, 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 seconds (80,37 minutes). 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 III. In particular, the indicators description and the network over which it is calculated are

detailed.

Table III. 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

The results from factor analysis using the set of variables selected by the genetic algorithm are detailed in Table IV. Usually, a number of factors equal to the number of eigenvalues higher than 1 is selected [17]. Consequently, up to 6 latent factors can be distinguished as result of factor analysis.

Table IV. 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
8	,741	2,964	83,401
9	,636	2,545	85,946
10	,532	2,127	88,073
11	,497	1,990	90,063
12	,429	1,716	91,779
13	,386	1,545	93,324
14	,330	1,318	94,642
15	,279	1,116	95,758
16	,249	,995	96,753
17	,193	,772	97,524
18	,172	,689	98,213
19	,143	,570	98,784
20	,124	,496	99,280
21	,078	,311	99,591
22	,045	,182	99,773
23	,032	,129	99,902
24	,017	,069	99,971
25	,007	,029	100,000

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.

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. An analysis of variance (ANOVA) has been performed to check the null hypothesis of equal population means. These null hypotheses have been rejected in all the cases with a significance value below 0,05. Using the information of the factor loadings as well as the mean values of the categorized groups of Universities, the following websites structure patterns can be highlighted (Table V):

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.

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.

Factor 3 refers to an egocentric structure, where the global network could be considered as the sum of more or less independent subnetworks.

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.

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

Finally, factor 6 represents web site with a structure dominated by one subnetwork, containing the most relevant information.

Table V. 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

Basically, the identified profiles of web site structures respond to two basic strategies when deciding their final structure [22]. 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, 3 and 4 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 [22]. This is the strategy of profiles identified by factors 5 and 6. Finally, factor 2 could be considered as a mixture of both strategies.

V. 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. 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).

References

[1] L. Björneborn and P. Ingwersen, "Toward a basic framework for webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 14, pp. 1216–27, 2004.

[2] M. Thelwall, "Bibliometrics to webometrics", *Journal of Information Science*, Vol. 34, no. 4, pp. 605-621, 2008.

[3] M. Thelwall, *Link Analysis: An Information Science Approach*, Amsterdam, Elsevier 2004.

[4] H.W. Park & M. Thelwall, "Hyperlink analysis: Between networks and indicators", Journal of Computer-Mediated Communication, Vol. 8, no. 4, 2003. http://www.ascusc.org/jcmc/vol8/issue4/park.html.

[5] S. L. Toral, M. R: Martínez Torres, F. Barrero, "Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis", *Information and Software Technology*, Vol. 52, Iss. 3, pp. 296-303, 2010.
[6] D. Iacobucci, *Graphs and matrices*. In Wasserman, S., & Faust, K. (Eds,), Social network analysis -- methods and applications. New York, NY: Cambridge University Press, 1994, pp. 92-166.
[7] J. L. Ortega, I. F. Aguillo, "Visualization of the Nordic academic web: Link analysis using social network tools", *Information Processing and Management*, Vol. 44, pp. 1624–1633, 2008.
[8] J. L. Ortega, I. F. Aguillo, "Mapping world-class universities on the web", *Information Processing and Management*, Vol. 45, pp. 272–279, 2009.
[9] R. Baeza-Yates, & C. Castillo, "Characterization of national web domains", *ACM Transactions on Internet Technology*, Vol. 7, Iss.2, pp. 1-32, 2007.
[10] K. Berlt, E. Silva de Moura, A. Carvalho, M. Cristo, N. Ziviani, T. Couto, "Modeling the web as a hypergraph to compute page reputation", *Information Systems*, In Press.
[11] G. Almpanidis, C. Kotropoulo, I. Pitas, "Combining text and link analysis for focused crawling—An application for vertical search engines", Information Systems, Vol. 32, pp. 886–908, 2007.
[12] W. Nooy, A. Mrvar, and V. Batagelj, *Exploratory Network Analysis with Pajek*, Cambridge University Press, New York, 2005.
[13] B. Yang, J. Qin, "Data collection system for link analysis", *Third International Conference on Digital Information Management*, ICDIM 2008, pp. 247-252.
[14] S. L. Toral, M. R. Martínez-Torres, F. Barrero, "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, 2009, pp. 405-419.
[15] S. L. Toral, M. R. Martínez-Torres, F. Barrero, and F. Cortés, "An empirical study of the driving forces behind online communities", *Internet Research*, Vol. 19, no. 4, pp. 378-392, 2009.
[16] C. Faba-Pérez, F. Zapico-Alonso, V. P. Guerrero-Bote, & F. de Moya-Anegón, "Comparative analysis of webometric measurements in thematic environments", *Journal of the American Society for Information Science and Technology*, Vol. 56, no. 8, pp. 779–785, 2005.
[17] A.C. Rencher, *Methods of Multivariate Analysis*. 2nd ed. Wiley Series in Probability and Statistics, John Wiley & Sons, 2002
[18] S. L. Toral, M. R. Martínez Torres, "International Comparison of R&D Investment By European, US and Japanese Companies", *International Journal of Technology Management*, Vol. 49, no. 1/2/3, pp. 107-122, 2009.
[19] J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975.
[20] M.R. Martínez-Torres, S.L. Toral, "Strategic group identification using evolutionary computation", *Expert Systems with Applications*, in press, 2010.
[21] D. A. Goldberg, *Genetic Algorithm – in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989.
[22] G. W. Tan, K. K. Wei, "An empirical study of Web browsing behaviour: Towards an effective Website design", *Electronic Commerce Research and Applications*, Vol. 5, pp. 261–271, 2006.