

Methods for Analysing Web Citations: A Study of Web-Coupling in a Closed Environment

CRISTINA FABÁ-PÉREZ, VICENTE P. GUERRERO-BOTE, AND FÉLIX DE MOYA-ANEGÓN

University of Extremadura, Library and Information Science Faculty, Badajoz, Spain and University of Granada, Library and Information Science Faculty, Granada, Spain

To reveal the structure of the relationships that establish themselves on the World Wide Web, one needs to apply tools that faithfully represent the virtual environment. Some of the most interesting relationships are those that are brought to light by web-coupling (the Web analogue of bibliographic coupling). We here propose an analysis of this type based on the common links that are generated within a

closed web environment, using multivariate statistics (principal component analysis, and multidimensional scaling) and a connection-based technique (Kohonen's self-organizing maps). The results show that it is possible to use the common links of web spaces in order to reveal the structures and the underlying relationships in a thematic closed environment.

Introduction

The citation analysis of scientific publications is a complex process that studies the relationship between citations/references (Price 1970) and the parties involved – the citer and the citee. Leydesdorff (1998) establishes a clear difference between the concepts of “citation” and “citation analysis”, associating them with practice and reflexive theory, respectively. This distinction notwithstanding, citation networks overall constitute a fundamental instrument with which to study the consumption of scientific information and to detect the authors, articles, and journals that have had the greatest impact on the scientific community (López-López 1996). One technique that uses citation analysis is bibliographic coupling. This was first popularised by Kessler (1963) and subsequently further developed by Vladutz and Cook (1984). The idea

itself, however, is due to Fano (1956), cited by Egghe and Rousseau (2002). Kessler introduced the term in order to emphasize that when two documents have at least one reference in common there exists a bibliographic bond between them. The underlying idea is that sharing common references is a sign of thematic proximity between documents, and that the degree of proximity will depend on the number of references in common. Kessler (1965) noted that the method had been named “bibliographic coupling” because it originated in the hypothesis that the reference lists of technical articles constitute a way by which the author indicates the intellectual environment in which he or she is operating and, if two documents present similar reference lists, there is an implicit relationship between them. This type of analysis then allows one to classify the research lines under study (Persson 1994).

Vicente P. Guerrero-Bote, (Corresponding author). University of Extremadura, Library and Information Science Faculty, Alcazaba de Badajoz (Antiguo Hospital Militar), 06071 Badajoz, Spain. Tel.: +3424259910. Fax: +3424286401. E-mail: vicente@alcazaba.unex.es

Cristina Fabá-Pérez, University of Extremadura, Library and Information Science Faculty, Alcazaba de Badajoz (Antiguo Hospital Militar), 06071 Badajoz, Spain. E-mail: cfabper@alcazaba.unex.es

Félix De Moya-Anegón, University of Granada, Library and Information Science Faculty, Campus Cartuja, Colegio Máximo, Granada, Spain. E-mail: felix@goliat.ugr.es

There have been numerous studies carried out to determine whether it is possible to apply traditional citation analysis to the environment of the Internet – to the World Wide Web in particular – and thus reveal the structure of the relationships that are established between the “Web spaces” – an expression introduced by Smith (1999) to refer to top-level domains, low-level domains, and groups of directories. To this end, they take the inlinks (links into a given Web space that is different from that of the source) and the outlinks (links out of a given Web space into another Web space different from that of the source) of Web spaces to be the respective analogues of citations-to and references in traditional scientific publications. This analogy has, in most cases, met with a favourable opinion in the literature (McKiernan 1996; Larson 1996; Almind and Ingwersen 1997; Boudourides, Sigrít and Alevizos 1999; Vreeland 2000; Björneborn and Ingwersen 2001; Cronin 2001; Chu, He and Thelwall 2002). Some authors consider the analogy to be risky, however, arguing that: (a) the dynamic and distributed nature of the Web permits a case – the two-way character of links – that is impossible in the analysis of traditional citations (Egghe 2000); (b) electronic links possess a set of intrinsic characteristics that differentiate them from scientific citations (Kim 2000); and (c) this new type of hypertext citation is too anonymous and superficial a system to merit the rank of scientific citation, and should rather be considered as simply another tool offered by the Web of the same level of importance as an e-mail or an electronic fax (Raan 2001). Some workers, such as Rousseau (1997), take an intermediate standpoint, considering that the study of “sitations” on the Web (“WebSiting”) although conceptually equivalent to that of traditional citations, has a slightly different sense because the motivation behind links and citations is not the same. The term “situation” was advanced by McKiernan (1996) and used by Rousseau (1997) to designate the relation between sites on the Internet. We follow Rousseau in the present work, using “situation” to differentiate the concept from that of citation in scientific publication. In this sense, although there is a diversity of motivations behind scientific citations, the process is accepted as being one of scientific persuasion (Brooks 1985). Electronic links, however, represent an aspect of “scientific-social-technological” behaviour (Kim

2000). Authors such as Borgman and Furner (2002) therefore consider it more appropriate to use the expression “link analysis” since its meaning is far broader than that of citation / reference and it covers all the possible modes of citation that could occur on the Web.

Some work on “WebSiting” or “link analysis” has focused specifically on what we call “Web-coupling”, i.e., the study of the relationships and frequencies of Web spaces that have outlinks in common. García-Santiago (2001), for instance, investigates the topology of the information on the World Wide Web by applying bibliometric techniques to display the topological organization of a national hypertext network. The basis of the study is to use the common references produced by Web links as a criterion of the proximity of the initial Web spaces (the origins). It starts from the hypothesis that the set of links that exist on the Web contain sufficient information to reveal the degree of relationship between Web spaces. According to the author, there are the following advantages in using the “Web-coupling” technique instead of cositation analysis (the study of Web spaces which are linked to conjointly by other Web spaces): (a) the upper limit on the number of common references is the number of outlinks in the origin Web space, i.e., the upper limit to the number of common links an origin can have is the number of all of its links, while there can be as many cositations as there are “sitations” in the entire network concerned; and (b) there is a clearer graduation of the numerical information in the common references as the numbers involved are far smaller than in the cositations.

The objective of the present work is to provide evidence in support of the hypothesis that it is possible to study the behaviour of the Web spaces in a closed thematic environment (represented by the case of the Spanish Region of Extremadura) on the basis of the common links that are generated within that environment. The analytical process that was followed lies within the framework of Web data mining, using techniques of multivariate statistics – principal component analysis (PCA) and multidimensional scaling (MDS) – and the topological organization generated by Kohonen’s self-organizing maps (SOMs) to determine the structure of the relationships between the Web spaces.

Material and method

We carried out the study on a population of Web spaces whose connection is their relationship with the Region of Extremadura. We located this population by looking for a source that gave us a set of Web spaces specializing in Extremadura from which we could retrieve new Web spaces on Extremadura by following the outlinks, finally generating a database of 1180 elements. Taking as the principal evaluation criterion the “authority” of the source, we selected “Extremadura on Internet” (<http://www.juntaex.es/todoWeb>), the Web server of the Junta of Extremadura (the supreme official organism in the Autonomous Community) which compiles Web sites, Web pages (including personal pages), and sets of Web pages lodged on other servers; in their retrieval, we considered the three categories conjointly under the generic term of “Web spaces” – of and about the Region of Extremadura.

Applying Web-coupling to this population of Web spaces, we found 894 of them with common links, and generated a Web-coupling matrix whose elements are the number of common links of the corresponding two Web spaces. This raw data matrix was subjected to a PCA to reduce the dimension of the problem and allow MDS to be applied. We characterized and labelled the factors according to their associated Web spaces, and represented the relationships amongst the factors using MDS and the relationships between the factors and the Extremadura Web spaces by means of SOMs. Examples of the application of the techniques of multivariate statistics – factorial analysis and MDS – to discovering the relationships among Web spaces are to be found in the work of Larson (1996), Chen and Cooper (2001), He and Hiu (2002) and Larsen et al. (2002).

Principal Component Analysis (PCA)

Factorial analysis techniques use a strategy of “informational parsimony” that allows one to identify a small number of factors that explain most of the variance observed in a greater number of variables (Herrero-Solana 2000). PCA is one of the most widely used techniques; its basic premise is that the best way to represent the linear relationship between two variables is by means of a straight-line regression. The PCA mechanism can be used to reduce pairs of variables to fewer dimensions in

order to simplify the graphical representation of the elements of the matrix. Each of these dimensions is known as a factor. These are ordered from the most important (the first factor) to the least important (the last factor). Generally the first factors (with the greatest eigenvalues – relative sizes or weights of the corresponding factor) account for a very high proportion of the variance, which means in practice that by themselves they can characterize the behaviour of the n-dimensional space. The remainder of the factors generally account for very little variance, so that they may be automatically dropped without the risk of losing much information. In the present case, 231 factors were found in the Web-coupling matrix, explaining a total variance of 83.67%. We then selected only the factors that by themselves explained more than 1% of the variance – the useful factors (García-Santiago 2001; Faba-Pérez 2003) – that together explained 40.62% of the variance.

In Web-coupling, applying a PCA, one will obtain a loading (calculated during the PCA, and indicating the component of each factor corresponding to each Web space), and the factors can then be characterized or labelled according to which Web spaces contribute the greatest loading in each case. Normally, Web spaces that do not surpass a certain threshold are not considered in characterizing the factor. In the present analysis, we only considered loadings greater than 0.7 to be significant (García-Santiago 2001; Faba-Pérez 2003).

Multidimensional Scaling (MDS)

In order to extract more information from the factors found by the Web-coupling PCA, we determined the factor correlation matrix (a symmetric matrix with dimension equal to the number of useful factors) from the matrix of loadings of the Web spaces in the useful factors. We then applied the MDS technique to that matrix.

One uses MDS to identify the dimensions that best show the similarities and distances between variables. Applications implementing some MDS calculation procedures use a similarity or distance matrix as input and calculate the coordinates in a two- or three-dimensional space that give similarities or distances as close as possible to those of the input matrix. To ensure that one has attained the best possible fit between the two sets of distances, one uses a statistical measure named

Table I. Web-coupling factors with variance > 1%.

Factor	Eigenvalue	Variance (%)	Accumulated Variance
1	136.3	15.23	15.23
2	41.64	4.66	19.89
3	40.31	4.51	24.40
4	35.5	3.97	28.38
5	22.4	2.53	30.91
6	20.5	2.29	33.20
7	17.0	1.90	35.11
8	15.9	1.78	36.89
9	12.0	1.35	38.24
10	11.5	1.29	39.53
11	9.76	1.09	40.62

Table II. Web-coupling characterization of useful factors.

Factor	Characteristics
F1	Institutional Web sites
F2	Culture and tourism
F3	Links to portal sites on tourism, business, and leisure
F4	Junta of Extremadura
F5	University of Extremadura
F6	Localities in Badajoz
F7	Entities
F8	Web site with maps
F9	Personal pages on the same server
F10	Sites with the same provider
F11	Trujillo

“stress” which represents that degree of fit between the observed and calculated similarities (Herrero-Solana 2000).

Self-Organizing Map of Kohonen (SOM)

We also used the topological organization generated by a Kohonen SOM to analyse the relationships between the Web spaces of our population and the factors found in the Web-coupling PCA, taking into account that the SOM, unlike the MDS, better represents proximity relationships than structural relationships. This type of neural network (Kohonen 1982, 1989, 1990, 1995; Kohonen et al. 1999) has often been used for text data mining (Lagus et al. 1999), and in particular to generate topological maps of a set of documents, including labelling the zones of influence of each word or term (Guerrero-Bote 1997; Lin 1997; Chen et al. 1998; Moya-Anegón, Jiménez-Contreras and Moneda-Corrochano 1998; Moya-Anegón et al. 1999; Kaski 1999; Lagus and Kaski 1999; Guerrero-Bote and Moya-Anegón 2001; Guerrero-Bote, Moya-Anegón and Herrero-Solana 2002a, 2002b).

The most characteristic feature of this type of neural network is a competitive layer (a layer in

which only a single neuron is activated by the presence of an input) which classifies (clusters) the inputs (on the basis of which neuron is activated by each input). The main difference with other competitive layers is that one simulates an influence of each neuron on its neighbours, which decreases with increasing distance from them. This has a biological basis, as has been found for certain primates (Hilera and Martínez 1995). As a consequence, a bubble of activity is formed in the layer by all those units which are close to the winner. These neighbouring units participate in the corresponding reinforcement of the learning. A result of this in the Kohonen self-organizing maps (unlike other competitive layers) is that units which are physically close participate conjointly in many cases of reinforcement of the learning and consequently respond to input vectors which are equally close. These neurons are usually arranged in a two-dimensional array. Since at times the only thing of interest is the topological organization performed by this layer, the whole set of vectors is selected for training only to see the resulting topological organization.

The algorithm which the neurons put into practice may be summarized in the following steps: (a) select as winning node (represented by a weight vector with the same dimension as the input) that closest to the presented vector; and (b) adjust the weight vectors of the winning node and of those corresponding to its neighbourhood by shifting them towards the input vector. Step (b) is only performed during the training stage, and in some cases the reinforcement is the same for the whole neighbourhood, while in others it decreases as the distance to the winner increases.

Results and discussion

The application of the PCA to our (894 x 894) Web-coupling matrix gave 231 factors explaining 83.67% of the variance. Table I lists the 11 factors that we considered useful (those that explain a variance of at least 1%). Their accumulated percentage of explained variance was 40.62%.

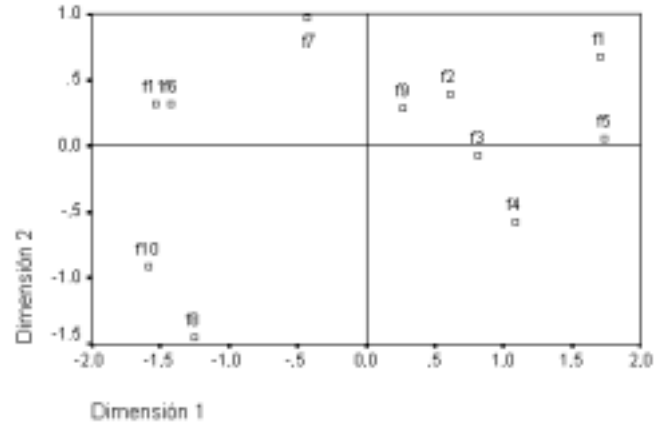
As we indicated above, in characterizing the useful factors we only considered Web spaces with loadings greater than 0.7. Table II lists the labels which in a summarized form characterize each of the 11 factors with variance greater than 1%.

One observes that PCA is a type of analysis that brings out many facets of the data. For instance, factors 6 “Localities in Badajoz” and 11 “Trujillo” represent different aspects of the “local environment”. Likewise, in factors 1 “Institutional Web sites” and 7 “Entities” there prevail the characteristics of the institutional Web site as against the personal pages (factor 9 “Personal pages”). Finally, factors 2 “Culture and tourism” and 3 “Links to portal sites on tourism, business, and leisure” reflect the importance of tourism in the Region. The discovery of these properties indicates that three principal characteristics form the backbone of the Web-coupling of Extremadura Web spaces: (1) their mainly local scope; (2) their institutional nature; and (3) the tourism economic sector.

We applied the MDS technique to the correlation matrix of the 11 factors. The results are shown in Figure 1, and correspond to a stress of 0.11. In the interpretation, we followed the four basic elements proposed by White and Griffith (1981):

- *Centre / Periphery.* For White and Griffith, in the MDS of an author cocitation analysis (ACA), the most important authors of a discipline are located in the centre of the representation. In the present case, however, there is no factor occupying the centre, but the tendency is rather to a general peripheral distribution over the map.
- *Clustering.* The overall view provided by this representation is of a greater clustering of factors located on the right-hand side of the figure, with the first 5 factors of the Web-coupling PCA, i.e., those explaining the greatest variances (factors 1 “Institutional Web sites”, 2 “Culture and tourism”, 3 “Links to portal sites on tourism, business, and leisure”, 4 “Junta of Extremadura”, and 5 “University of Extremadura”), and factor 9 “Personal pages”. On the left-hand side of the map, one observes two clusters of just two elements each: factors whose Web spaces with greatest loadings represent some local type of information (factors 6 “Localities in Badajoz” and 11 “Trujillo”), and factors whose Web spaces with greatest loadings are lodged on the same providers (factors 8 “Web site with maps” – many of these sites are on www.geocities.com – and 10 “Sites with the same provider – www.gratisWeb.com”). These clusters indicate that factors near each other have similar loading patterns or, which comes to the same thing, patterns of links in common with the factors in the same group.
- *Identification of the axes.* The horizontal dimension is distributed symmetrically along the corresponding axis. It runs from factors 10 “Sites with the same provider – www.gratisWeb.com” and 11 “Trujillo”, past the factors characterized by localities in Extremadura (factors 6 “Localities in Badajoz” and 8 “Web site with maps”), past factor 7 “Entities”, past personal pages and culture / tourism (factors 9 “Personal pages”, 2 “Culture and tourism”, and 3 “Links to portal sites on tourism, business, and leisure”), and past factor 4 “Junta of Extremadura”, to end at the factors labelled as institutional (factors 1 “Institutional Web sites” and 5 “University of Extremadura”). One could say that in a certain sense this dimension is related to the location / breadth of Web spaces. The second dimension is not symmetric, and runs from factor 8 “Web site with maps”, past factor 10 which is very distant from the rest (except for factor 8), past factors 4, 3, and 5, and past factors 11, 6, 9, and 2 (these last factors are practically at the same height), and past factor 1, to end at factor 7.

Figure 1. Web-coupling MDS (two-dimensional space).



ism”, and 3 “Links to portal sites on tourism, business, and leisure”), and past factor 4 “Junta of Extremadura”, to end at the factors labelled as institutional (factors 1 “Institutional Web sites” and 5 “University of Extremadura”). One could say that in a certain sense this dimension is related to the location / breadth of Web spaces. The second dimension is not symmetric, and runs from factor 8 “Web site with maps”, past factor 10 which is very distant from the rest (except for factor 8), past factors 4, 3, and 5, and past factors 11, 6, 9, and 2 (these last factors are practically at the same height), and past factor 1, to end at factor 7.

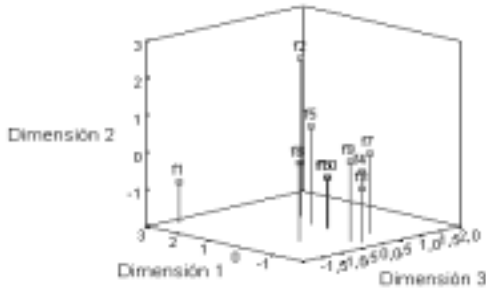
- *Relationships between nearby elements.* Factors 11 “Trujillo” and 6 “Localities in Badajoz” merit special attention because of their closeness (although they are distant from the rest). This indicates that these factors labelled by their local character have loading patterns that are very similar to each other.

Figure 2 is a three-dimensional plot of the 11 Web-coupling factors. While the 3D representation is more complex, it does offer a little more fidelity by reducing the stress (0.09). The correct interpretation of this type of structure requires all the possible orientations of the model to be displayed, not just a single facet captured on paper. Although the drop lines for each point give an idea of the spatial distribution of the elements, the true capacity of the representation is revealed when it is analysed through a dynamic, virtual reality type of interface.

To obtain the topological organization of the SOM, we used a neural network with a hexagonal topology of 30 x 50 neurons that we presented with the Web spaces represented by 11-dimensional vectors whose components corresponded to the loadings in each of the useful factors.

Figure 3 shows a representation of the topology of the network after the training procedure. Each

Figure 2. Web-coupling MDS (three-dimensional space).



hexagon with a dot inside represents a neuron. Training was carried out in two phases: (i) 1000 cycles, 0.05 learning rate, and 40 for the initial neighbourhood parameter progressively reduced to 1 over the course of the 1000 cycles; (ii) 10 000 cycles, 0.02 learning rate, and 10 for the initial neighbourhood parameter again progressively reduced to 1 over the course of the 10 000 cycles.

Eleven unit vectors were generated corresponding to the 11 factors (all the components of these vectors are set to zero except that corresponding to the given factor which is set equal to unity). These vectors were used, on the one hand, to determine the winning neuron for each factor, i.e., the neuron whose weight vector is closest to the factor's unit vector (as is shown in the figure), and, on the other, to find the factor closest to each neuron's weight vector so that the set of all neurons corresponding in this sense to a given factor determines the factor's zone of influence, represented in Figure 3 by separation lines.

Thus, by means of the network we were able to classify all the Extremadura Web spaces that were linked to onto the nodes of a hexagonal grid. In this type of representation, the shading represents the distances between the weight vectors of neighbouring neurons, i.e., between the centroids of the clusters to which the said neurons give rise. Thus, the distances, and therefore the colours, are an indication of the closeness of the Web spaces (classified in the corresponding neurons) as determined by the factors. To use a geographical metaphor, the light colours indicate valleys in which cities (neurons) with large populations are better communicated and hence more closely related, while the dark colours indicate mountain ranges that constitute zones of isolation with small scattered populations (Kohonen et al. 1999).

To improve the legibility of Figure 3, we have used each factor's key (F1, F2, etc.) instead of its

label. Keys with the larger font size correspond to factors that have zones of influence and whose winning neuron is in that zone (the key tag is placed on that winning neuron). The smaller font refers to zones of influence of factors whose winning neuron is outside that zone. One observes that all the factors have a zone of influence in which are classified the Web spaces whose greatest loading is in that factor. Also, most of them have an associated valley in which (using the above metaphor) the cities (neurons) are better communicated and therefore more closely related. The existence of a dominant zone for all the factors means that there are Web spaces whose principal factor is the one indicated and that there are no others that are more relevant. Factors 2 "Culture and tourism" and 3 "Links to portal sites on tourism, business, and leisure" have two zones of influence: one closer to its winning neuron and another zone without a winning neuron. A detailed analysis of the two factors shows that, in factor 2, the Web spaces in general and those of greatest loadings in particular are characterized by culture and tourism, but that there also exists another small group of Web spaces with a certain tendency towards institutional Web sites of a business nature. This last group extends towards the zone with no winning neuron that borders factors 4 "Junta of Extremadura" and 7 "Entities" (related in part to the institutional nature of their Web spaces and to business themes). With respect to factor 3, its Web spaces in general and those of greatest loadings in the factor are characterized by linking to an Extremadura portal site related to tourism, business, and leisure. But there also exists a group of Web spaces without links to these portal sites but with information on localities and institutions of Extremadura. This last group seems to extend towards the zone with no winning neuron that borders factors 6 "Localities in Badajoz" and 7 "Entities".

Some factors have an intermediate result, with a zone of influence located on a plateau between the factors that they are most closely related to (an example is the case of factor 10 "Sites with the same provider - www.gratisWeb.com"). The large valleys are mainly associated with the factors that explained the most variance (factors 1 "Institutional Web sites", 2 "Culture and tourism", 3 "Links to portal sites on tourism, business, and leisure", and 4 "Junta of Extremadura"), although factor 7 "Entities" is an exception.

Figure 3. Classification of Web-coupling factor vectors (hexagonal neighbourhood), labeling the factor domains, the winning neuron of each factor, and the winning neuron of the Web spaces (first 200 of the ranking).

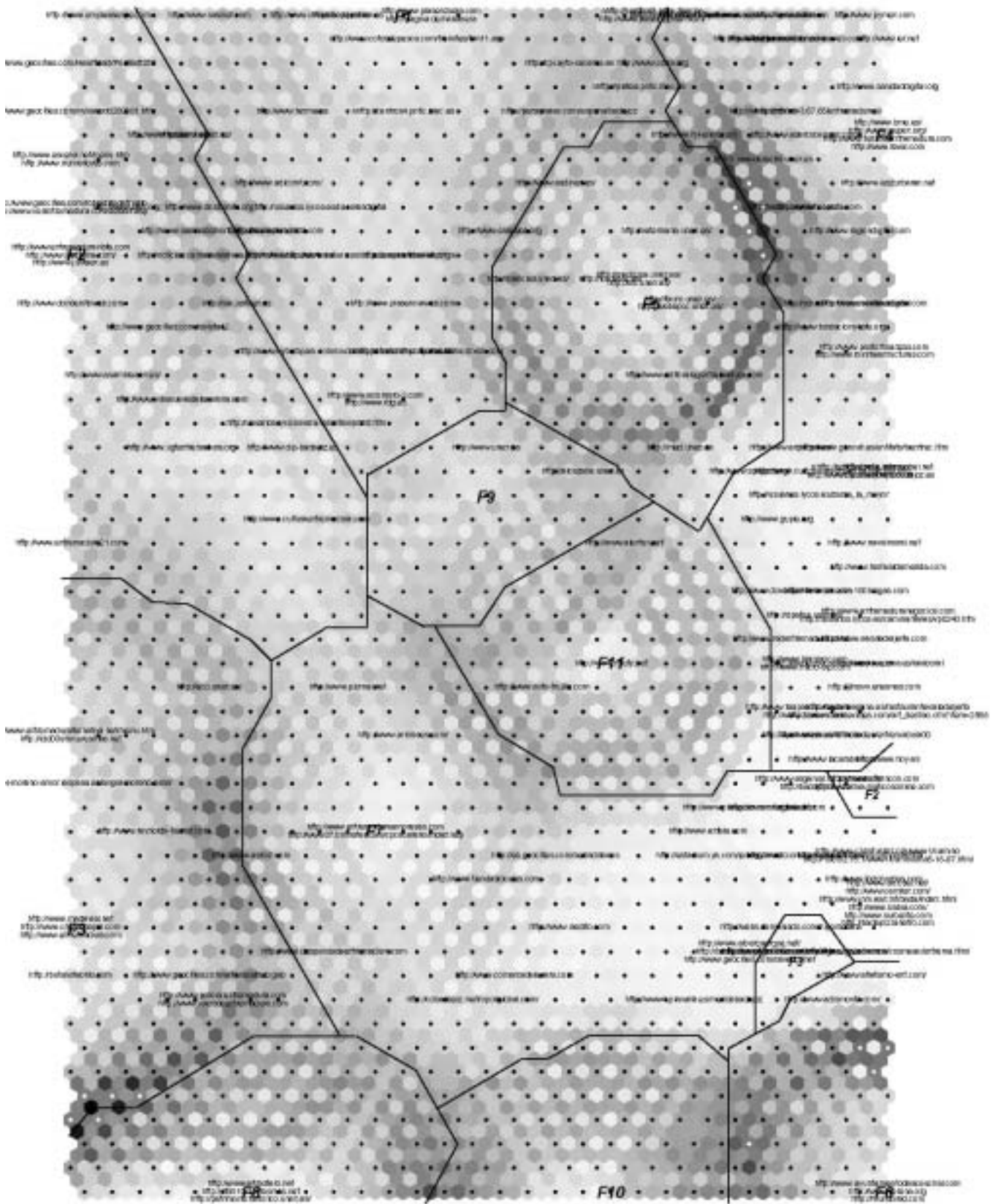
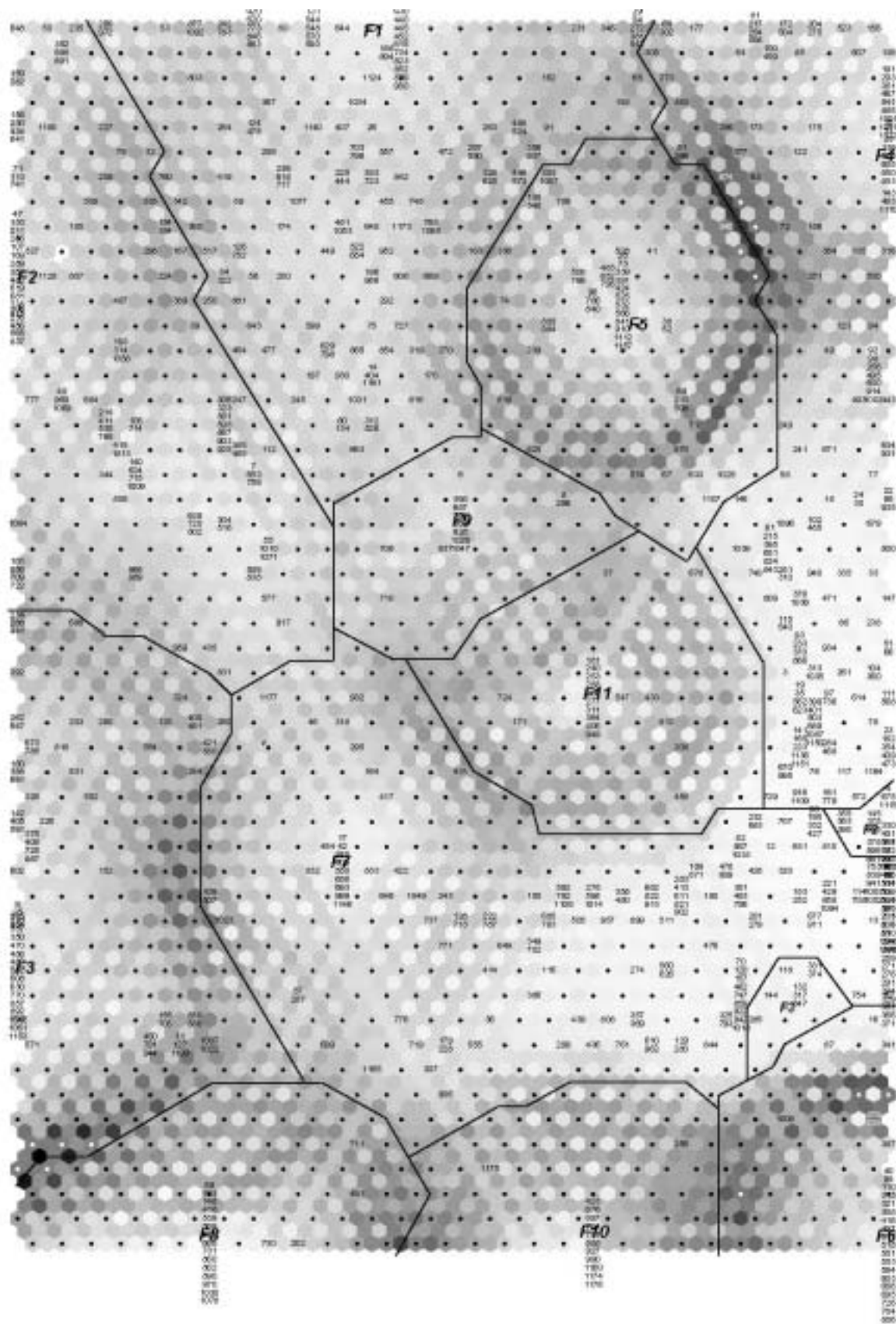


Figure 4. Classification of Web-coupling factor vectors (hexagonal neighbourhood), labeling the factor domains, the winning neuron of each factor, and the winning neuron of the Web spaces (with the rank).



Focusing on the neighbour relationships of the map, i.e., on the adjacency between factors, one observes few dark areas in general, implying that there is little distance between neighbouring domains, except between factors 4 “Junta of Extremadura” and 5 “University of Extremadura”, and between factors 3 “Links to portal sites on tourism, business, and leisure” and 8 “Web site with maps” where there are mountainous zones on the border between the domains. This indicates that the Web spaces classified in these domains, and that therefore have the greatest loading in these factors, share very few links in common.

Figure 3 shows certain interesting relationships that the MDS was unable to pick up. The zones of influence of the first four factors (F1–F4) lie on the periphery of the map. That of factor 7 “Entities” occupies one of the central positions and borders eight of the other ten possible factors (factors 4 “Junta of Extremadura”, 11 “Trujillo”, 9 “Personal pages”, 2 “Culture and tourism”, 3 “Links to portal sites on tourism, business, and leisure”, 8 “Web site with maps”, 10 “Sites with the same provider – www.gratisWeb.com”, and 6 “Localities in Badajoz”). It also neighbours factors 2 and 3 in their respective zones with no winning neuron, and presents valleys and plateaus on all its borders. This indicates that the Web spaces classified in this domain also have loadings in all those other domains, whereas in the MDS representation factor 7 appeared isolated by having a Web-space loading pattern different from the rest.

Likewise, the relationship of closest proximity that found with MDS, which was of a markedly local character (factors 11 “Trujillo” and 6 “Localities in Badajoz”), has now disappeared in the SOM. Other clusters that were not very close to each other in the MDS representation, such as factors 8 “Web site with maps” and 10 “Sites with the same provider – www.gratisWeb.com”, or factors 1 “Institutional Web sites” and 5 “University of Extremadura”, which are of an institutional but not necessarily local character, still border each other.

This type of network also allows one to place the Web spaces of our population on the grid together with the factors. In the case of Figure 3, the Web spaces shown are the top 200 ranked by a measure of quality (calculated as a weighted sum of formal indicators of quality plus a logarithmic size factor (Faba-Pérez, 2003)). They are principally distributed between the factors associ-

ated with the aforementioned large valleys, i.e., factors 1, 2, 3, 4, 5, and 7. In contrast, however, these Web spaces with the classification’s greatest weights are either absent from the factors that explained the least variance (factor 10 “Sites with the same provider”) or have a negligible presence (factor 11 “Trujillo”). In the case of factor 9 “Personal pages”, while this includes only two of the first 200 Web spaces of our classification, these are two of the highest ranked (6 and 8).

Figure 4 shows the same Kohonen network as in the previous figure but now with the 1180 Web spaces of the population represented by their rank according to the aforementioned quality measure. The figure shows that the factors that account for most Web spaces are those associated with the largest valleys (with a particularly marked grouping in factors 4 “Junta of Extremadura” and 7 “Entities”). Nevertheless, other factors which practically included no Web spaces of the top-ranked 200 of the classification now appear with an acceptable number of Web spaces.

Conclusions

We have confirmed that, with the techniques applied in the present work, it is possible to reveal the relationships that are established between the Web spaces of a closed thematic environment through the study of their shared links (Web-coupling). With the PCA, we were able to find the principal factors involved in the Web-coupling of Extremadura Web spaces, with the MDS to visualize the relationships between these factors, and with the SOM to organize topologically the Region’s Web spaces including the labelling of each factor’s winning zone.

These relationships were determined on the basis of the links contained in the pages of the Web spaces themselves. In analogy with the case of Bibliographic Coupling, this leads us to speak of the view that Web spaces – or their authors – have of the rest of the elements – their world – with which they are linked. This is different from the case of cositation / cocitation studies, in which the relationships are determined on the basis of the cositations / cocitations made by the rest of the elements, and then one speaks of the view of each element held by the rest of the elements.

Hence, the picture of the behaviour of the Extremadura Web spaces given by the MDS and

SOM representations led us to draw three main conclusions:

1. In general, the overall view that the Extremadura Web spaces have of the world of the Web varies little from one to another – a mark of their closed character. This conclusion arises from an observation of the Web-coupling SOM, in which the dominant light colours are those indicative of communication between factors, indicating that the relationships or patterns of common links of Web spaces of different domains are similar.
2. It is notable in this SOM representation that the greatest degree of proximity is found between clusters of an institutional nature (official and local). This can again be interpreted as these Extremadura Web spaces having very similar images of the Web.
3. In the case of the MDS representation, the Extremadura Web spaces of a local character (which are very numerous, since the environment is markedly rural) are extremely close to each other. This behaviour indicates that the Extremadura Web spaces that are local in nature have very similar views of the World Wide Web environment, as is reflected in their shared links. This is, in a certain manner, a logical finding since these Web spaces form a pre-eminently closed (local) environment.

Acknowledgments

This work was financed by the Junta de Extremadura – Consejería de Educación Ciencia & Tecnología – and the European Social Fund, as part of the “Programa de Ayudas para la Realización de Proyectos de Aplicación de las Tecnologías de la Información y la Comunicación” (published in D.O.E. nº 120, 17 October 2000).

References

- Almind, T. C. and Ingwersen, P. 1997. Informetric analyses on the World Wide Web: methodological approaches to ‘Webometrics’. *Journal of Documentation* 53(4): 404–26.
- Björneborn, L. and Ingwersen, P. 2001. Perspectives of Webometrics. *Scientometrics* 50(1): 65–82.
- Borgman, C. L. and Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review and Information Science and Technology (ARIST)* 36: 3–72.
- Boudourides, M. A., Sigrit, B. and Alevizos, P. D. 1999. *Webometrics and the self-organization of the European Information Society*. URL: <http://hyperion.math.upatras.gr/Webometrics> [viewed October 16, 2000].
- Brooks, T. A. 1985. Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science* 36(4): 223–29.
- Chen, C. et al. 1998. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science* 49(7): 582–603.
- Chen, H. and Cooper, M. D. 2001. Using clustering techniques to detect usage patterns in a Web-based Information System. *Journal of the American Society for Information Science & Technology* 52(11): 888–904.
- Chu, H., He, S. and Thelwall, M. 2002. Library and Information Science Schools in Canada and USA: a Webometric perspective. *Journal of Education for Library and Information Science* 43(2): 110–25.
- Cronin, B. 2001. Bibliometrics and beyond: some thoughts on Web-based citation analysis. *Journal of Information Science* 27(1): 1–7.
- Egghe, L. 2000. New informetric aspects of the Internet: some reflections – many problems. *Journal of Information Science* 26(5): 329–35.
- Egghe, L. and Rousseau, R. 2002. Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics* 55(3): 349–61.
- Faba-Pérez, C. 2003. *Análisis cibernético de la información WEB: el caso de Extremadura en Internet*. PhD. Thesis, University of Granada, Spain.
- Fano, R. M. 1956. Information theory and the retrieval of recorded information. In: Shera, J. M., Kent, A. and Perry, J. W., eds. *Documentation in Action*. New York: Reinhold Publ. Co.: 238–44.
- García-Santiago, M. D. 2001. *Topología de la información en la World Wide Web: modelo experimental y bibliométrico en una red hipertextual nacional*. PhD. Thesis, University of Granada, Spain.
- Guerrero-Bote, V. P. 1997. *Redes Neuronales aplicadas a las Técnicas de Recuperación Documental*, PhD. Thesis, University of Granada, Spain.
- Guerrero-Bote, V. P. and Moya-Anegón, F. de. 2001. Reduction of the dimension of a document space using the fuzzified output of a Kohonen network. *Journal of the American Society for Information Science* 52: 1234–41.
- Guerrero-Bote, V. P., Moya-Anegón, F. de and Herrero-Solana, V. 2002a. Document organization using Kohonen’s algorithm. *Information Processing & Management* 38: 79–89.
- Guerrero-Bote, V. P., Moya-Anegón, F. de and Herrero-Solana, V. 2002b. Automatic extraction of relationships between terms by means of Kohonen’s algorithm. *Library & Information Science Research* 24: 235–50.
- He, Y. and Hiu, S. C. 2002. Mining a Web Citation Database for author co-citation analysis. *Information Processing and Management* 38: 491–508.
- Herrero-Solana, V. 2000. *Modelos de representación visual de la información bibliográfica: aproximaciones multivariantes y conexionistas*. PhD. Thesis, University of Granada, Spain.
- Hilera, J. R. and Martínez, V. J. 1995. *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. Madrid: RAMA.

- Kaski, S. 1999. Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*. London: Institution of Electrical Engineers: 940–45.
- Kessler, M. M. 1963. Bibliographic coupling between scientific papers. *American Documentation* 14:10–25.
- Kessler, M. M. 1965. Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation* 16(3): 223–33.
- Kim, H. J. 2000. Motivations for hyperlinking in scholarly electronic articles: a qualitative study. *Journal of the American Society for Information Science* 51(10): 887–99.
- Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43(1): 59–69.
- Kohonen, T. 1989. *Self-Organization and Associative Memory*. Berlin: Springer Verlag.
- Kohonen, T. 1990. The Self-Organizing Map. In: *Proceedings of the IEEE*: 1464–80.
- Kohonen, T. 1995. *Self-Organization Maps*. Berlin, Heidelberg: Springer Verlag.
- Kohonen, T. et al. 1999. Self-organization of a massive text document collection. In: Oja, E. and Kaski, S., eds. *Kohonen Maps*. Amsterdam: Elsevier: 171–82.
- Lagus, K. and Kaski, S. 1999. Keyword selection method for characterizing text document maps. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*. London: Institution of Electrical Engineers: 371–76.
- Lagus, K. et al. 1999. WEBSON for textual data mining. *Artificial Intelligence Review* 13(5–6):345–64.
- Larsen, J. et al. 2002. Webmining learning from the World Wide Web. *Computational Statistics & Data Analysis* 38(4): 517–32.
- Larson, R. R. 1996. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace". In: Hardin, S., ed. *Proceedings of the 59th Annual Meeting of the American Society for Information Science (Baltimore, Maryland, 1996)*, Information Today, Medford, New Jersey: 71–78. URL: <http://sherlock.berkeley.edu/asis96/asis96.html> [viewed October 14, 2000].
- Leydesdorff, L. 1998. Theories of citation? *Scientometrics* 43(1): 5–25.
- Lin, X. 1997. Maps displays for Information Retrieval. *Journal of the American Society for Information Science* 48(1): 40–54.
- López-López, P. 1996. *Introducción a la Bibliometría*. Valencia: Promolibro.
- McKiernan, G. 1996. *CitedSites (sm): Citation Indexing of Web Resources*. URL: <http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm> [viewed February 24, 2000].
- Moya-Anegón, F. de et al. 1999. NeuroISOC: un modelo de red neuronal para la representación del conocimiento. In López-Huertas, M. J. and Fernández-Molina, J. C., eds. *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información. Actas del IV Congreso ISKO-España (EOCONSID'99)*. Granada: ISKO-España: 151–156.
- Moya-Anegón, F. de, Jiménez-Contreras, E. and Moreda-Corrochano, M. de la. 1998. Research fronts in library and information science in Spain (1985–1994). *Scientometrics* 42(2): 229–46.
- Persson, O. 1994. The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science* 45(1): 31–38.
- Price, D. J. de Solla. 1970. Citation measures of hard science, soft science, technology and non-science. In: Nelson, C. C. and Pollock, D. E., eds. *Communication among scientists and engineers*. Lexington, Mass.: D. C. Heath and Co.: 3–22.
- Raan, A. F. J. van. 2001. Bibliometrics and Internet: some observations and expectations. *Scientometrics* 50(1): 59–63.
- Rousseau, R. 1997. Sitations: an exploratory study. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics* 1. URL: <http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html> [viewed September 5, 2000].
- Smith, A. G. 1999. *The impact of Web sites: a comparison between Australasia and Latin America*. URL: <http://www.vuw.ac.nz/~agsmith/publns/austlat/> [viewed May 14, 2001].
- Vladutz, G. and Cook, J. 1984. Bibliographic coupling and subject relatedness. Challenges to an Information Society. In: *Proceedings of the 47th ASIS Annual Meeting*: 204–207.
- Vreeland, R. C. 2000. Law libraries in hyperspace: a citation analysis of World Wide Web sites. *Law Library Journal* 92(1): 9–25.
- White, H. D. and Griffith, B. C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science* 32(3): 163–71.

Editorial history:

paper received 13 May 2003;

final version received 16 December 2003;

accepted 29 December 2003.