

Methods for the Analysis of the Uses of Scientific Information: The Case of the University of Extremadura (1996–7)

VICENTE P. GUERRERO-BOTE, MARÍA J. REYES-BARRAGÁN, FÉLIX DE MOYA-ANEGÓN,
VICTOR HERRERO-SOLANA

Facultad de Biblioteconomía y Documentación, Universidad de Extremadura, Badajoz, Facultad de
Biblioteconomía y Documentación, Universidad de Granada, Campus Cartuja, Granada, Spain

The present study is an institutional domain analysis from the perspective of the information requirements of the scientific-technical area of the University of Extremadura in order to optimise access to and availability of scientific publications. The data were the international publications of the University, their authors and the departments to which they belonged and their reference lists obtained from the SCI (ISI

Science Citation Index ®). The results are presented and a methodological analysis is made using various statistical techniques (cluster analysis, factorial analysis, and multidimensional scaling) to determine the structure of the departmental relationships in the institution, and the outputs of Kohonen neural networks to display the relationships between journals and departments.

Introduction

Specialist journals are currently among the most important basic tools for scientists for the transmission of knowledge, as well as for the generation of new knowledge to be transmitted in its turn. Their high cost makes the optimisation of their availability and accessibility a key to the development of scientific knowledge. The University of Extremadura is in the process of optimising the whole complex of its library system. One part of the process is the study of the information requirements of its researchers. A domain analysis (Hjørland & Albrechtsen 1995) has been performed on the different fields of research. Here we present the data corresponding to the scientific-technical area.

According to the University Statutes in force, the departments are responsible for the organiza-

tion and development of the research and teaching in their respective areas of knowledge in one or various centres of the University. Part of this responsibility is the administration of resources dedicated to the acquisition of publications. It seems natural therefore to base the study on these units rather than on the faculties, technical colleges, or university schools responsible for the administrative management and organization of the lecture courses. It has to be kept in mind that the departments of the University of Extremadura are academic units grouping together sets of researchers. For the purpose of optimisation of management, the departments have a minimum number of researchers. In practice, this has obliged researchers from distinct scientific fields to group together into a single department when the number of researchers in each field did not attain this minimum level (Morphological Sciences

Vicente P. Guerrero-Bote. Facultad de Biblioteconomía y Documentación; (Antiguo Hospital Militar); Universidad de Extremadura, E - 06071 Badajoz, Spain. Tel.: (924) 25.99.10-15. Fax: (924) 25.99.57. E-mail: vicente@alcazaba.unex.es.

María J. Reyes-Barragán. Facultad de Biblioteconomía y Documentación, Universidad de Extremadura, 06071 Badajoz, Spain. E-mail: mjreyes@alcazaba.unex.es

Félix de Moya-Anegón, E-mail: felix@goliat.ugr.es; Victor Herrero-Solana, E-mail: victorhs@ugr.es; Facultad de Biblioteconomía y Documentación, Universidad de Granada, Campus Cartuja, Granada, Spain

and Cellular and Animal Biology). This must be taken into account in interpreting the results.

The characterization of the departments is perhaps a complex matter subject to a certain level of controversy. It is a necessary step, however, in setting up the proposed basic collection and in resolving in some form the issue of its location. This is no longer a matter of delimiting the information requirements of a group of scientists, but goes beyond this in how to determine the information consumption measured in uses of the various scientific publications by the departments to which the researchers belong.

With respect to this issue, we are aware of the difficulties involved in quantifying the uses of collections, and in particular of serial publications, as is reflected in the literature on the topic. Hayes (1981) poses the question of what is a "library use". Is it when a certain journal is taken down from the stack and browsed, or when it is taken over to the table? In either case, its measurement would require the collaboration of librarian and of users, and the adoption of a uniform criterion throughout the system. As we here have a departmental structure, however, where there exist small "libraries" with no librarian or qualified personnel to manage them, the use of this type of method is not feasible.

It seems to us that the important thing is to detect which journals are being used and by whom. The former will be used in setting the acquisition policy and the latter to determine the location.

Following Peat (1981), who criticizes the most commonly used procedures, and taking into consideration what we mentioned above, that these publications are the basic medium of transmission and generation of knowledge, we feel that the best way to study the demand for information is on the basis of the knowledge generated. We propose as the method to determine the real use of the collection the analysis of the bibliographic references which appear in the institution's own publications. We shall use these references as indicators of the most used journals in the researchers' investigations. We add who uses these journals within a departmental structure so as to detect the uses that a journal is put to with respect to the departments.

Citation analysis has been used to disentangle the structure of the various fields of science and to determine with greater precision the develop-

ment of scientific activity since the 1960s, when Price (1965) uncovered the relationship between the network formed by scientific works (linked by their citations and references) and the structure of a given scientific field. Techniques of data analysis based on multivariate statistics allow the dimensions to be reduced and conclusions drawn.

There are two main types of data on which these techniques are based: bibliographic coupling and co-citation. The analysis of common references was first evaluated by Kessler (1963) and later developed by Vladutz and Cook (1984). It has now been implemented in the ISI databases (*SCI-CD*®, *SSCI*® and *A&H*®). The underlying idea is the study of the citing works on the basis of the shared references.

Co-citation, initially presented by Small (1973) and Marshakova (1973) and later developed by Small & Sweeney (1985), is based on the concept that the co-citation between two works (the number of co-occurrences in the reference lists of scientific articles) is indicative of their topical affinity. This technique has also been used for the study of authors and journals (White and Griffith 1981; White 1981, 1983; White and McCain 1997, 1998; Leydesdorff 1987; Leydesdorff and Cozzens 1993; Van der Besselaar and Leydesdorff 1997; Persson 1994; Moya-Anegón et al. 1998a) as well as for the study of articles.

Occasionally the two types of data have been combined (Persson 1994). Co-citation study analyses the intellectual basis, while bibliographic coupling classifies the lines of research under study.

In the study of the structure and composition of scientific fields, the use of co-citation has predominated over bibliographic coupling. The reason is mainly that the former characterizes an article, author, or journal on the basis of the similarities and the utility that the rest of the community finds that item by referencing it together with others. This does not depend on the will of the authors. The basis of bibliographic coupling, however, is formed by the authors' own judgments.

In the present case, we wish to study an institution. The source therefore is the publications of its researchers. From these, we chose their references, because our aim was to analyse information requirements, and the sources used may be considered indicative of the researchers' informa-

tion requirements when they are carrying out their work.

Data

We used the SCI bibliographic database to give the count of citations to scientific journals in the area of science and technology. Searches were set up in the *address* field (address of the authors belonging to the University of Extremadura). We thus obtained the output of the University of Extremadura in the period consisting of the years 1996 and 1997, a time interval that we considered to be sufficient to determine the information requirements of the researchers.

A priori no problem is posed by the biases in the database itself, which collects almost exclusively a corpus of English language journals, since English-speaking countries are leaders in the ambit of science. No matter whatever the case is, however, one currently has no choice but to use this database.

To characterize the departments, we first generated a file of authors of articles in science and technology who belonged to the University, and assigned them their corresponding department. The exhaustive and unique identification of the authors who produced in this area was a costly problem to solve in terms of effort. The reason was mainly the lack of normalization in the names of the authors obtained from the SCI database. This was especially so for those who sign with two family names, and accentuated when these family names were themselves very common. This situation was provoked in part by the authors themselves who did not always sign or refer to themselves in a consistent way. There is also a lack of coherence on the part of the creators of the database who apply arbitrarily the set norms for Spanish or for English-speaking authors. The problem was solved by reference to the University's directory, to data provided by the Vice-rectorate for Research, and with the aid of experts in each field.

After the identification of the signing authors, the references of each article were assigned to the corresponding departments, weighted by the number of authors belonging to those departments. Imagine, for instance, that we had a scientific article signed by four authors (A, B, C, D). Authors A and B belong to department 1, C to de-

partment 2, and D does not belong to the University. The uses of the journals cited in this article will be weighted by two for department 1, by one for department 2, and author D will be eliminated.

This design to determine the journal use by researchers of the different departmental units could also have been approached by setting up a search for the corresponding departments in the address field. The problem then would have been the delimitation of signing authors belonging to different departments. We opted for the first possibility noted, as we considered it to be faster and more objective, with the bonus effect of eliminating any possible mistakes due to normalization in the addresses.

All calculations were computerized, based on four linked lists corresponding to authors, departments, articles, and journals, and using programs to extract the data from the input records retrieved from the SCI. Some of the data had to be completed manually, as was the case for the department to which each author belonged, the impact factor of the journals, and their availability in the University of Extremadura's libraries.

Given this data structure, it was possible to automate all the computations needed for the aforementioned bibliometric study, as well as those required for the present study. For the latter case, we generated the vectors that would later be processed by means of Kohonen's algorithm, and a matrix of journal use whose columns represented the journals and rows the departments, with each element indicating the number of uses of each journal by each department.

By way of summary of the resulting data, 474 SCI records were retrieved, involving 21 departments and 2467 cited journals.

Method

The traditional interpretation of multidimensional matrices such as that generated here has been by way of applying multivariate statistical techniques. We wish also to propose the use of Kohonen's neural network, which has been applied by White et al. (1998). We therefore first applied cluster analysis, multidimensional scaling, and factorial analysis (from amongst the statistical methods), and then followed by contrasting these results with those from the Kohonen network.

We applied the three statistical methods to the departments. To this end, we pre-processed the use matrix by finding the scalar product of the vectors corresponding to the departments. The result is a square matrix with as many rows and columns as departments. Each element now corresponds to the scalar product of the vectors of the corresponding departments, i.e., it is indicative of coincidences of journal use between departments. This is similar to the pre-processing to use co-citation instead of citations. It accentuates similarities and differences, and thus allows a better interpretation of the results. The diagonal elements, which correspond to the scalar product of each department with itself, for the departments with many publications and therefore much use of the journals led to great distances between them and the rest of the departments, even with the departments they have most in common with. For this reason, as in other cases, they were replaced with the maximum of the rest of the elements of that department (i.e. of the scalar products of that department with the rest). Given that the number of articles, and therefore the number of references, varied greatly between the departments, we used a function that eliminated these differences of scale transforming the scalar product matrix into a Pearson correlation matrix.

Although we shall not discuss the first three techniques of multivariate statistics, which are very well known and much used in similar studies, we shall comment briefly on Kohonen's algorithm given the novelty of its application in this field (Kohonen, 1982, 1989, 1990, 1995; Kohonen et al., 1999). We used this algorithm to analyse the journals cited by each department. The most characteristic feature of this type of network is a competitive layer that classifies (clusters) the training inputs. The main difference with other competitive layers is that each neuron exerts an influence 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 of the hidden layer which are physically close respond to input vectors

which are equally close. These neurons are usually arranged in a two-dimensional array. For this reason, at times the only thing which is of interest is the clustering performed by the hidden layer, and the whole set of training vectors is selected only to see the resulting topological organization.

The algorithm which the neurons put into practice may be summarized in the following steps:

- Select as winning node (represented by a weight vector with the same dimension as the input) that closest to the presented vector.
- Adjust the weight vectors of the winning node and of those corresponding to its neighbourhood by shifting them towards the input vector (in some cases the reinforcement is the same for the whole neighbourhood, and in others it decreases as the distance to the winner increases).

This type of network has recently been used for *textual data mining* (Lagus et al. 1999), and in particular to generate topological maps of a set of documents, even labelling the zones of influence of each word or term (Moya-Anegón et al. 1998b, 1999; Chen et al. 1998; Guerrero-Bote 1997; Guerrero-Bote & Moya-Anegón 2001; Guerrero-Bote et al. 2002; Lin 1997; Kohonen et al. 1999; Kaski 1999; Lagus and Kaski 1999).

In our case we shall apply to it as input certain patterns that represent the different journals. To this end, we used a weighting scheme similar to the IDF used in the vector space model (Salton and McGill 1983). This has also been found to have a good behaviour when used with this type of network (Lin 1997; Guerrero-Bote 1997; Guerrero-Bote & Moya-Anegón 2001; Guerrero-Bote et al. 2002; Chen et al. 1998). To represent a journal, we use as many components as we have departments. The weights are assigned as:

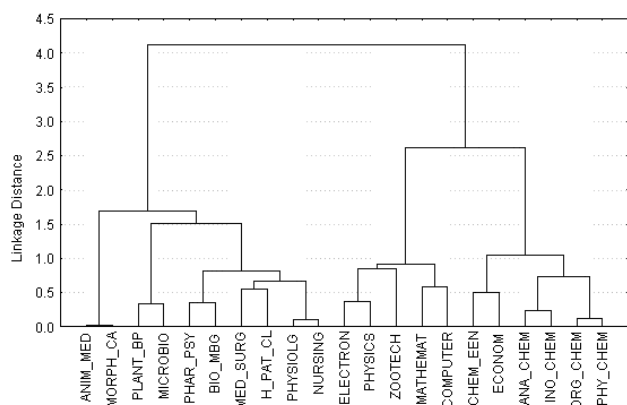
$$a_{ij} = d_{ij} \cdot \log \frac{P}{p_j}$$

where:

- a_{ij} = weight assigned to department d_i in journal P_j .
- d_{ij} = number of uses by department d_i of journal P_j .
- p_j = number of journals used by department d_i .
- P = total number of journals.

In this way greater weight is given to the uses by those departments using fewer journals, since these uses characterize the department more.

Figure 1: Hierarchical clustering of the departments (Ward's method as the clustering rule, and the complement to unity of the correlation as distance).



Results

Figure 1 depicts the result of applying to the aforementioned matrix of correlations between the departments a hierarchical clustering algorithm using Ward's method as the clustering rule, and the complement to unity of the correlation as distance. One observes in the dendrogram that there exist two main groups with a great relationship distance (greater than 4). The first one could label as "the life sciences" and the second as "the hard sciences" (although, as will be seen below there exist certain exceptions).

In the *Life sciences* group, the first cluster is of the pair formed by the departments of *Animal Medicine* and *Health and Morphological Sciences and Cellular and Animal Biology*. These are very close to each other, and separated from the rest by a distance of 1.7. Then another pair splits off formed by the departments of *Plant Biology and Production* and *Microbiology* at a distance of 1.5. The other departments are more closely related (at a distance less than 0.8) and form a subgroup closer to the study of *human life and health*.

The *hard sciences* group is divided into two major clusters (separated by a distance of 2.6). One is mainly related to Chemistry and the other to Physics and Mathematics. Here one can already appreciate, however, some of the aforementioned exceptions: the department of *Zootechnics* appears in the *Mathematics* and *Physics* related cluster, and *Applied Economics and Business Management* in the Chemistry group. This is something that these

Table I: Principal component analysis (PCA) of the correlation matrix. The six eigenvalues greater than unity were extracted. Those loadings greater than 0.3 (after performing a *varimax* orthogonal rotation for the departments to have the greatest weights in few factors) were kept.

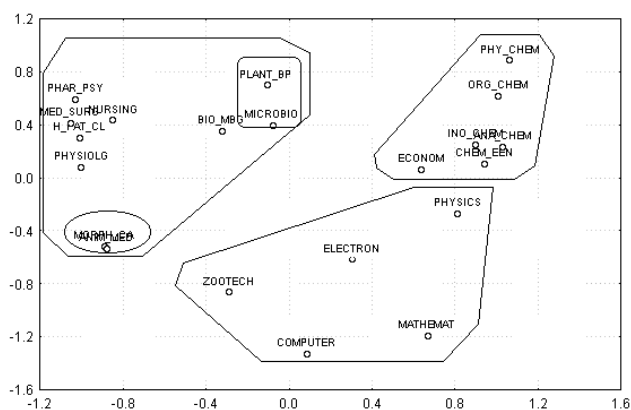
FACTOR (Eigenvalue)	1 (4.88)	2 (3.85)	3 (2.65)	4 (1.93)	5 (1.63)	6 (1.17)
<i>Applied Economics and Business Management</i>	0.81					
<i>Inorganic Chemistry</i>	0.71					0.38
<i>Chemical and Energy Engineering</i>	0.71					
<i>Nursing</i>		0.86				
<i>Physiology</i>		0.85		0.33		
<i>Human Pathology and Clinics</i>		0.73				
<i>Medical/Surgical Specialities</i>		0.70				
<i>Biochemistry and Molecular Biology and Genetics</i>		0.66			0.58	
<i>Mathematics</i>			0.73			
<i>Electronics and Electromechanical Engineering</i>	0.30		0.73			
<i>Physics</i>	0.59		0.70			
<i>Computer Science</i>			0.63	0.32		
<i>Animal Medicine and Health</i>				0.97		
<i>Morphological Sciences and Cellular and Animal Biology</i>				0.94		
<i>Plant Biology and Production</i>					0.76	
<i>Zootechnics</i>			0.45		0.67	
<i>Microbiology</i>	0.49			0.36	0.66	
<i>Pharmacology and Psychiatry</i>		0.55			0.56	
<i>Physical Chemistry</i>						0.93
<i>Organic Chemistry</i>	0.34					0.91
<i>Analytical Chemistry and Electrochemistry</i>	0.50					0.76

statistical techniques detect, and whose explanation we shall defer until later.

We used 21 components in the correlation matrix to represent each department. (These components are indicative of the information requirements in common with the rest of the departments.) Using factorial analysis, we can discover whether there exist a smaller number of hidden factors that determine each component's value. Each department will have a certain loading for each factor. One may then use those having the greatest loadings in a factor to represent that factor (usually setting the threshold at 0.7). Likewise, the factors with the greatest loadings in a department can be used to represent that department. These factors could be interpreted as groups of information requirements, with the loadings indicating what is required in each department.

Table I lists the results of the principal component analysis (PCA). Six eigenvalues greater than unity were extracted (given in parentheses below the number of the factor), which gave rise to the corresponding factors (for those less familiar with this type of statistical analysis, each factor has an associated eigenvalue, and this divided by the number of variables, 21 in our case, is an indication of the proportion of the variance which that factor explains). We kept only those loadings greater than 0.3 (after performing a *varimax* or-

Figure 2: Multidimensional scaling of the departments using journal use coincidences to represent the departments (stress of 0.14).



thogonal rotation for the departments to have the greatest weights in few factors), and then ranked the departments according to their factor of greatest loading.

One deduces from the table that the first and sixth factors correspond to the Chemistry cluster (revealed in the previous analysis). Nevertheless, another three departments acquire a significant weight in factor one.

The second factor corresponds to the study of *human life and health*. The departments classified previously in the same group acquire a significant loading in this factor. It is notable, however, that some departments have significant weights in other factors. This is the case with *Physiology* in factor 4, which is led by the departments of *Animal Medicine and Health* and *Morphological Sciences and Cellular and Animal Biology* (which, in turn, form the first group to be split off from that of the life sciences). Likewise, the departments of *Biochemistry and Molecular Biology and Genetics* and *Pharmacology and Psychiatry* have a very significant weight in factor 5, which is led by the department of *Plant Biology and Production*. Thus, these three factors characterize the departments corresponding to the life sciences.

Amongst those surpassing 0.7 in factor 3 are *Mathematics, Electronics and Electromechanical Engineering, and Physics*. Computer Science also has a significant weight here, as it does too in factor 4 because of the collaboration of one of its members with three of the department of *Morphological Sciences and Cellular and Animal Biology*. *Zootechnics*, which was classified in the same group by the cluster analysis, has a significant

weight in this factor and even more so in factor 5. Nonetheless, it is classified in this group by the previous method.

Perhaps the most confusing case is *Microbiology* which has significant weights in three factors, although it ended as classified into the factor that had the highest weight.

The result of the multidimensional scaling technique is depicted in Figure 2, with a stress of 0.14. This value is satisfactory even for Boyce et al. (1994). This figure is a two-dimensional representation of the departments, aimed at assimilating the distances existing in the 21-dimensional space. We have marked on the figure the main classes determined by the clustering algorithm. One sees the confirmation of the information obtained from the previous analyses, with the different clusters logically being situated around the periphery. One also sees how the two anomalously classified departments are, in their turn, far from the centres of their corresponding clusters. *Zootechnics* clearly tends towards the *Life sciences*. *Applied Economics and Business Management* tends to distance itself from its cluster, though without approaching any other in particular.

If we analyse the horizontal dimension, there is a symmetric distribution along the corresponding axis and a flow from the departments mainly dedicated to the study of *human life and health*, through the rest of the *Life sciences, electronics and computing*, to reach *chemical sciences*. The second dimension starts from *Morphological Sciences and Cellular and Animal Biology* (the most exact sciences), evolving towards *Life sciences* on the one hand, and *chemical sciences* on the other. This dimension is not symmetric. The distances are greater in the mathematical part. Notable in this dimension is the closeness of *Zootechnics* to the mathematical extreme.

To look further into these anomalous relationships and try to explain them a little more, we performed the same study on the collaborations in the articles (instead of the coincidences in journal use). This led to a similar representation shown in Figure 3. One sees more relationships which are anomalous through collaboration, including the formation of clusters which disappear when the citations are taken into account in the journals. *Zootechnics* is still close to *Physics* (of the 13 articles of *Zootechnics*, 3 are in collaboration with *Physics* and 1 with *Pharmacology*

and Psychiatry). The cluster centred on *Applied Economics and Business Management* has its explanation in that 3 of its 7 articles are in collaboration with *Chemical and Energy Engineering* and another 3 with *Animal Medicine and Health*. In this case there also appear anomalously the department of *Morphological Sciences and Cellular and Animal Biology* (due to its only having collaborations with the departments of *Computer Science and Electronics and Electromechanical Engineering*) and that of *Biochemistry and Molecular Biology and Genetics* (due to its collaborations with *Physics*). The said collaborations have a great influence in the study. We found, however, that the inclusion of the references (which go deeper into the content of the work) achieves a more rational classification. It brings the department of *Chemical and Energy Engineering* closer to the rest of the Chemistry departments (although it does carry along with it *Applied Economics and Business Management*). Likewise, the departments of *Biochemistry and Molecular Biology and Genetics*, *Morphological Sciences and Cellular and Animal Biology*, and *Animal Medicine and Health* are brought closer to the *Life sciences*, although *Zootechnics* leaves definitively. It is also responsible for the formation of the cluster which gathers together the departments which study *human life and health*.

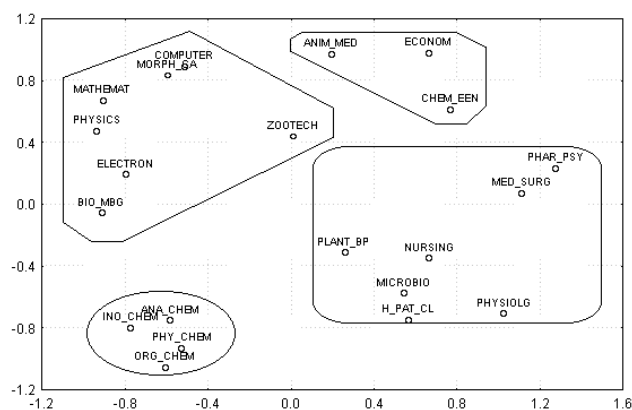
The University of Extremadura is relatively young, and is currently in a stage of growth. Some of its departments are therefore not yet sufficiently consolidated (due to their youth), as is detected in the study.

With this statistical analysis we were able to observe the similarities between the various departments as a function of their information requirements. But we were able to learn little about these requirements. We shall now present the results and potential of using Kohonen's neural network algorithm.

For this purpose, we start with the vectors that represent the publications, with 21 components as a function of the uses of the 21 departments. We shall use the weighting scheme commented on at the end of the previous section, and a network of 20 by 30 neurons in a hexagonal topology, so that each neuron has six neighbours.

Figure 4 shows the representation of the topology of the network after training. Each of the hexagons containing a dot represents a neuron. The shadings represent the distances between the

Figure 3: Multidimensional scaling of the departments using collaborations in the articles to represent the departments (stress of 0.14)

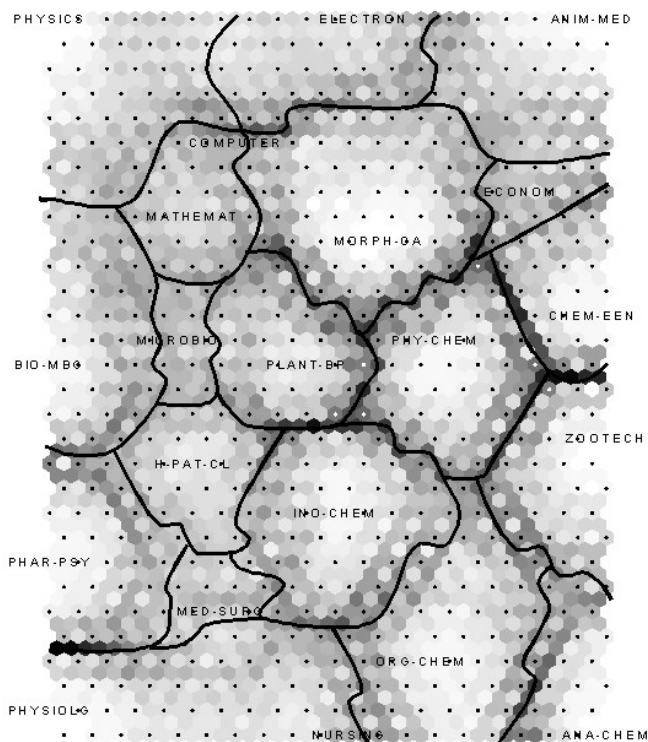


weight vectors of neighbouring neurons, i.e., between the centroids of the clusters each neuron gives rise to. We also generated 21 unitary vectors representing each department. These have all their components equal to zero except the one corresponding to the said department. These were applied firstly to determine the neuron with the closest weight vector, tagging itself to that neuron as is shown in the figure. Then they were applied to find the department which was closest to each neuron's weight vector, so that joining all the neurons which are closest to a department we have the zone of influence of that department (depicted in the figure by separation lines). This technique was used by Lin (1997), Chen et al. (1998), Moya-Anegón et al. (1998b, 1999) with terms and documents, and by Campanario (1995) with journals.

Put in other words, by means of the network we have managed to classify all the cited scientific publication on the nodes of a hexagonal grid. The distances, and hence the colours, are indicators of the topical proximity of the journals (classified in the corresponding neurons), bearing in mind that they are characterized by each department's uses. An easy way to interpret it is as a topological map, where the light colours indicate the valleys in which the cities (neurons) are better communicated and therefore more related. The darker colours indicate mountain ranges which are barriers of isolation.

Most of the departments have an associated valley and domain region, where the journals which are mainly used by that department are classified. There are two departments, *Computer Science*

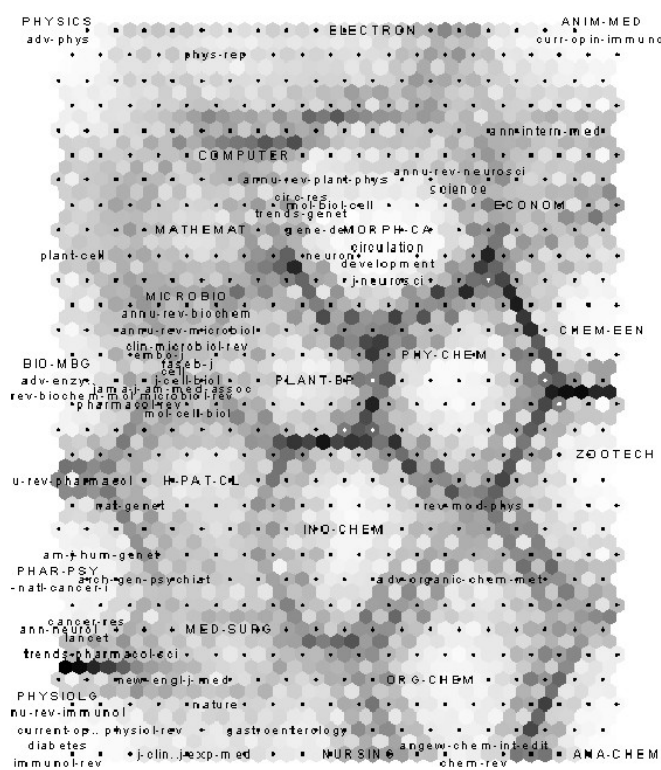
Figure 4: Representation of the topology of the network after training. Each of the hexagons containing a dot represents a neuron. The shadings represent the distances between the weight vectors of neighbouring neurons. We also generated 21 unitary vectors representing each department. These were used to determine the neuron with the closest weight vector, tagging itself to that neuron as is shown in the figure, and they were applied to find the zone of influence of that department (depicted in the figure by separation lines).



and *Nursing*, which do not have these domain zones. This is because they use journals that are mostly used by other departments. While neither department has participated in many articles (4 and 6, respectively), there are others such as *Plant Biology and Production* which, with a similar number (4), have established a domain zone and even a valley. The difference lies in that the journals used by this last department were rarely used by the others, and this is in turn a consequence to a great degree of there being little collaboration with researchers from other departments. The previous two departments, however, participated in work where most of the collaborators belonged to other departments. This implies that the network was able to detect some departments as yet unconsolidated in their research.

These departments appear on a dark crest at the end of the domain zone of a department with which they had most collaborated (as is the case

Figure 5: The distribution of the journals used that have an impact factor greater than 7.5 over the representation of the topology.



for *Nursing*), or even on a crest which borders the zones of the departments with which it had collaborated (*Physics, Morphological Sciences and Cellular and Animal Biology*, and *Mathematics* in the case of *Computer Science*). Others, such as *Applied Economics and Business Management*, obtain an intermediate result, a domain zone with no valley, located on a *meseta* between those departments with which it had collaborated most. The major valleys are associated with consolidated departments. This is the case of *Physics* with 92 articles in this two-year period, or *Physiology* which even has two associated valleys (two main lines of research), etc.

The overall topology of the network, however, is less clear than that of the multi-dimensional scaling method. But it should not be forgotten that the journals have now been classified.

With this type of network we may also situate journals on the grid together with the departments. Figure 5 shows the distribution of the journals used that have an impact factor greater than 7.5. This allows us to observe the quality of the journals used in each department.

Figure 6 depicts the distribution of the journals available in our University, so that one can see any possible gaps in the collection. (In this figure, we have only shown those journals which were cited more than nine times by these articles, to avoid their titles being overwritten.)

In these last two figures, in zones where there are a great many journals, there is a reduction in the legibility of their titles because they may overwrite one another. There exist different ways of obtaining the complete information, such as using shorter keys, or generating a list as a function of the grid co-ordinates. For the present work, however, we have shown a small number of journals, and have when necessary re-touched the figure slightly to improve legibility.

Conclusion

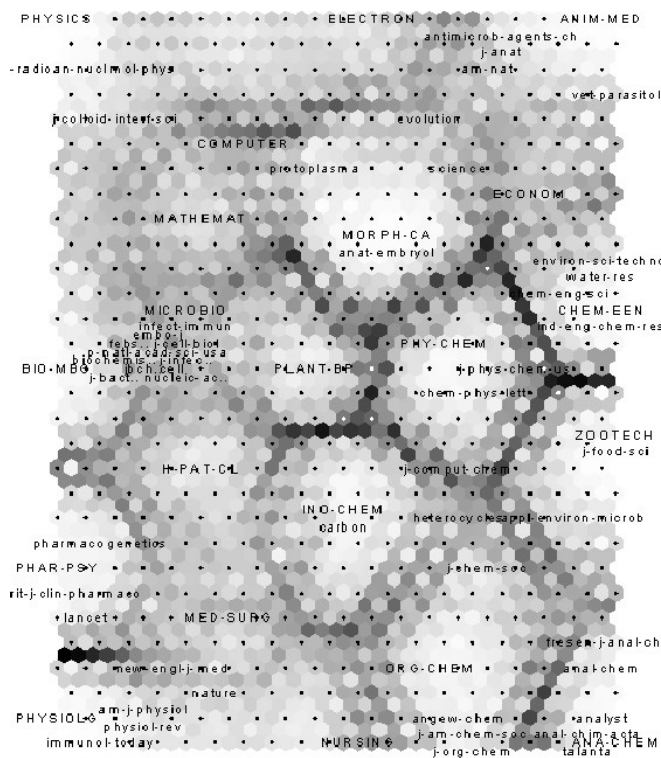
The objective of the present work was the analysis of the results and a test of the methodology. The latter can be divided into two parts: the use of *statistical methods* and the use of *Kohonen's neural network*.

In the first part, a novelty of the present work was the use of the same statistical methods employed on many occasions to study the front lines of research, the conceptual bases, or the relationships between journals of determined disciplines or fields of knowledge. Just as in these cases, it is of particular interest to disentangle the research structure of the institution under study. In the present case it has allowed us to see the relationships between the departments, to classify them, and so forth. We were able to verify the results that the techniques yielded.

These results indicated that the underlying research structure (of the scientific-technical area) is made up of two large groups. The first corresponds to the *life sciences*, is fairly heterogeneous and has various subgroups within it. The second corresponds to the *hard sciences*, and is in turn divided into two large subgroups, one of considerable coherence dedicated to *chemical sciences* (mainly), and the other less coherent (with the departments farther apart) which includes the *exact sciences*.

We were also able to observe that the said structure is largely due to collaboration between researchers of different departments (above all for the anomalous relationships). When the ref-

Figure 6: The same as Figure 5, but showing the periodicals available (those that one has access to) in our University.



erences were included, however, the structure became clearer, yielding another more logical structure.

We used the Kohonen maps to classify the journals, and in turn to place them relative to the different departments. The resulting topological organization allowed us to observe the relationships between departments, which largely coincided with the previous results (although of poorer quality as such). Nevertheless, with the study of the domain areas of each department, as well as of the distance between the neurons, we were able to study each department's zone of research and its relationship with the rest. Particularly interesting were the cases in which no domain area was found because the department was not consolidated, or the domain area contained no valley because of the dispersion of the works that had been published, or the area contained more than one valley because of the existence of various lines of research.

It was also possible to place on the resulting map each of the journals used. In our case, we only depicted those available in the University and those which surpassed a certain impact factor. One can set up any filter one likes, however,

and it is also possible to study the departments in which each journal is required.

Finally, as can be seen, the study reveals the utility and complementarity of the two techniques. The statistical techniques, using a representation for each department according to its information requirements, provide an overview of the relationships between them. They do not allow, however, one to study the relationship between the departments and the journals. The neural algorithm, allows a great many journals to be classified according to the departments that use their information together with a classification of the departments, so that one can see the relationships between the two separately or globally. It is known (Kohonen 1995) that this algorithm used to study the departments allows a better representation of their proximity to each other as well as of their zone of influence, while a better overall vision is obtained with such techniques as MDS (multidimensional scaling). It should also be emphasized that the network obtains this information by classifying a great number of journals related to the departments. If the representations of the departments are split up, the information that is provided will be far poorer.

References

- Boyce, B., Meadows, C. T., and Kraft, D. 1994. Measurement in information. London: Academic Press.
- Campanario, J. M. 1995. Using neural networks to study networks of scientific journals. *Scientometrics* 33(1): 23–40.
- Chen, H., Houston, A., Sewell, R., and Schatz, B. 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.
- Guerrero-Bote, V. 1997. Redes Neuronales aplicadas a las Técnicas de Recuperación Documental. PhD Thesis, Universidad de Granada, Spain.
- Guerrero-Bote, V. & Moya-Anegón, F. 2002. Reduction of the Dimension of a Document Space using the Fuzzified Output of a Kohonen Network. *Journal of the American Society for Information Science* (in press).
- Guerrero-Bote, V.; Moya-Anegón, F. & Herrero-Solana, V. 2002. Document organization using Kohonen's algorithm. *Information Processing & Management* 38(1): 79–89.
- Hayes, R. M. 1981. The distribution and use of library materials: analysis of data from the University of Pittsburgh. *Library Research* 3(3): 215–260.
- Hilera, J. R., and Martínez, V. J. 1995. Redes neuronales artificiales, fundamentos, modelos y aplicaciones. Madrid: RAMA.
- Hjørland, B., and Albrechtsen, H. 1995. Towards a new horizon in information science: domain-analysis. *Journal of the American Society for Information Science* 46(6): 400–425.
- Kaski, S. 1999. Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99). London: Institution of Electrical Engineers: 940–945.
- Kessler, M.M. 1963. Bibliographic coupling between scientific papers. *American Documentation* 14: 10–25.
- 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. *Proceedings of the IEEE* 78(9): 1464–1480.
- Kohonen, T. 1995. Self-organization Maps. Berlin, Heidelberg: Springer Verlag.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. 1999. Self organization of a massive text document collection. In Oja, E. and Kaski, S. ed. Kohonen Maps. Amsterdam: Elsevier: 171–182.
- Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. 1999. WEBSOM for textual data mining. *Artificial Intelligence Review* 13(5/6): 345–364.
- Lagus, K. and Kaski, S. 1999. Keyword selection method for characterizing text document maps. Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99). London: Institution of Electrical Engineers: 371–376.
- Leydesdorff, L. y Cozzens, SE. 1993. The delineation of specialities in terms of journals using dynamic journal set of the Science Citation Index. *Scientometrics* 26(1): 133–156.
- Leydesdorff, L. 1987. Various methods for the mapping of science. *Scientometrics* 11(5–6): 295–324.
- Lin, Xia. 1997. Maps Displays for Information Retrieval. *Journal of the American Society for Information Science* 48(1): 40–54.
- Marshakova, V. 1973. System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya, Series II* (6): 3–8.
- Moya-Anegón, F., Jiménez Contreras, E., De la Moreda Corrochano, M. 1998a. Research fronts in Library and Information Science in Spain (1985–1994). *Scientometrics* 42(2): 229–246.
- Moya-Anegón, F., Herrero-Solana, V., and Guerrero-Bote, V. 1998b. Virtual reality interface for accessing electronic information. *Library and Information Research News* 22(71): 34–39.

- Moya-Anegón, F., Moscoso, P., Olmeda, C., Ortiz-Repiso, V., Herrero-Solana, V., and Guerrero-Bote, V. 1999. NeuroISOC: un modelo de red neuronal para la representación del conocimiento. In López Huertas, MJ, and Fernández Molina, JC, eds. La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información. Actas del IV Congreso ISKO-España (EOCONSID'99). Granada: ISKO-España: 151–156.
- Peat, W. L. 1981. The use of research libraries: a comment about the Pittsburgh study and its critics. *Journal of Academic Librarianship* 7(4): 229–231.
- 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, J. D. De Solla. 1965. Networks of scientific papers. *Science* 149: 510–515.
- Salton, G. y McGill, M J. 1983. Introduction to modern information retrieval. New York: McGraw-Hill.
- Small, H. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24 (4): 265–269.
- Small, H., and Sweeney, E. 1985. Clustering the Science Citation Index using co-citations: 2 – mapping science. *Scientometrics* 8(5–6): 321–340.
- Van der Besselaar, P. y Leydesdorff, L. 1997. Mapping Change in Scientific Specialties: A Scientometric Reconstruction of the Development of Artificial Intelligence. *Journal of the American Society for Information Science* 47(6): 415–436.
- Vladutz , G., and Cook, J. 1984. Bibliographic coupling and subject relatedness." Challenges to an Information Society, *Proceedings of the 47th ASIS Annual Meeting* 21: 204–207.
- White, H. D., Griffith, B. C. 1981. Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science* 32(3): 163–171.
- White, H. D. 1981. Cocited author retrieval online: an experiment with the social indicators literature. *Journal of the American Society for Information Science* 32(1): 16–21.
- White, H. D. 1983. A cocitation of the social indicators movement. *Journal of the American Society for Information Science* 34(5): 307–312.
- White, H. D., McCain, K. W. 1997. Visualization of Literatures. In Williams, ME, ed. Annual Review of Information Science and Technology 32. Medford, NJ: Information Today: 99–168.
- White, H. D., McCain, K. W. 1998. Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science* 49(4): 327–355.
- White, H., Lin, X., and McCain, K. 1998. Two modes of automated domain analysis: multidimensional scaling vs. Kohonen feature mapping of information science authors. In Mustafa el Hadi, W., Maniez, J., & PollitErgon Verlag, S. Eds. Structures and relations in knowledge organization: Proceedings of the Fifth International ISKO Conference. Würzburg: Ergon Verlag: 15–29.

Editorial history:

Paper received 3 December 2001;

Final version received 19 February 2002;

Accepted 18 April 2002.