A STUDY OF THEMATIC AREAS IN ECONOMY BY A MEASURE OF SIMILARITIES BASED ON A KERNEL

FRANCISCO VELASCO, LUIS GONZÁLEZ-ABRIL, JUAN ANTONIO ORTEGA and JUAN ANTONIO ÁLVAREZ

SUMMARY

Some results on the similarities among thirteen subjects in the Economy area between 1990 and 2003 are presented and a virtual analysis of their interrelations is carried out. To this end, an intuitive interpretation of the similarity measure between two sets based on the Kernel method is defined, which provides a useful graphical representation. Several tables are considered that show the potentiality of this similarity index.

The first stage of research or innovation in any scientific research is having up-to-date knowledge of previous works on the subject. Nowadays there are many databases available for consultation and, thus, it is easy to obtain up-to-date information. An inexperienced researcher may have doubts when selecting a research field, and about choosing the most appropriate subject to achieve the best possible results. His supervisors can obviously help him, although if he can obtain up-to-date information about the interrelation between different subjects, and about how productive they are currently in the research world, he will have stronger reasons to choose one of them. This is true not only for a new researcher, but for all researchers in any subject, since it is always essential to know the thematic relations between distinct areas. One subject can be strongly related to another; that is, they can be similar and evolve together; or, they can be dissimilar in the sense of gradually becoming more distant from each other. It seems that this question can be solved by analyzing absolute values, such as for instance the number of works which are simultaneously referred to two particular subjects. But this localized view does not allow to see all the relations between every subject from a global perspective, and for this reason it is necessary to build a certain bounded measure that will enable to decide whether two subjects are close to each other or disconnected. It would be also helpful to know how this similarity has developed over time, and how it is expected to change in the near future.

This kind of problems have already been treated and have given rise to the development of a wide range of multivariate analysis tools and visual procedures that have been adapted to disciplines such as Sciencemetry (Callon *et al*., 1986; Coulter *et al*, 1998), Informetry (Egghe and Rousseau, 1990; Wolfram, 2000), Bibliometry (Noyons *et al*., 2002; Buter and Noyons, 2002; Mijac and Ryder, 2009) or Webometry (Almind and Ingwersen, 1997; Rousseau, 1997; Larson, 1996), in order to analyze the documental databases from a visual perspective. These methods are known as "procedures of dimension reduction" and their common characteristic is to transform the available information and store it in a two- or three-dimensional space where it is easier to analyze. These procedures of dimension reduction are used in all disciplines related to the treatment of scientific

KEYWORDS / Kernel / Scientific Production / Similarity / SSCI /

Received: 03/13/2009. Modified: 02/26/2010. Accepted: 03/01/2010.

Francisco Velasco Morente. Ph.D. in Mathematics, Faculty of Mathematics, University of Seville (US) Spain. Lecturer, Department of Applied Economics I, US. Spain Address: FCCEE, Department of Applied Economics I, Avenida Ramón y Cajal, 1, Sevilla, España. e-mail: velasco@us.es

Luis González-Abril. Ph.D. in Economics, US, Spain. Lecturer, Department of Applied Economics I, US. Spain. e-mail: luisgon@us.es

Juan Antonio Ortega Ramírez. Ph.D. in Computer Science, US, Spain. Lecturer, Department of Computing System and Languages, US, Spain. e-mail: jortega@us.es

Juan Antonio Álvarez García. Ph.D. in Computer Science, US, Spain. Lecturer, Department of Computing System and Languages, e-mail: jaalvarez@us.es

information and can be classified into neuronal and statistical procedures. The neuronal procedures are based on the learning capacity of the neuronal networks, such as Kohonen's net (Lin and Marchionini, 1991; Kohonen, 1998; Kohonen *et al*., 2000), in order to achieve a dimensional reduction for the bibliometric data, and the statistical procedures include clustering as one of these techniques. There is a wide variety of statistical tools used to reduce the bibliometric data dimension (Kinnucan *et al*., 1987), and among them we can highlight the multidimensional scaling (Deus, 2001; Klock and Buhman, 1999), whose development began in the psycho-psychic area.

The associated words method (Braam *et al*., 1991; Callon *et al*., 1991; Courtial, 1994; Grivel and Francois, 1995; Baños and Contreras, 1998; Coulter *et al*., 1998) is one of the best-known statistical techniques for establishing hierarchic order. This method is based on a graph where the key words are represented by knots and the arches refer to how frequently the related key words appear. From these graphs, such techniques can find and represent centers of interest concealed in the documents; that is, zones strongly related and consistent networks, susceptible of being interpreted as "hot points" or "attraction poles" of powerful informative intensity (Baños and Contreras, 1998; de la Rosa *et al*., 2005).

In contrast with the previous ones, a new procedure was described in González *et al*. (2005), based on a study of frequencies and having its foundations in the Statistical Learning Theory (Li *et al*., 2007; Srebro, 2007; Vapnik, 1998; Burrell, 2005; Clara, 2006; Zhang and Fu, 2006; Stentiford, 2007; González-Abril *et al*., 2009a, b). The measure of similarity described in González et al. (2005) is used to measure the connections between several subjects. Nevertheless, this measure has several drawbacks, because it depends on the weights of each set with respect to the rest. So, it does not provide a unified scale and does not allow to make comparisons. For this reason, a new similarity index is defined in this paper, which allows for the solving of these inconveniences.

This method could be applied to any field of knowledge. The reason why attention is driven to economics relies on the high concern that the problems stemming from a crisis produce, not only because of the current situation, but also because there has always been a tendency for crises to arise throughout history. In addition, it seems interesting to know which are the latest topics in the field of economics, and the relationships between them. Thus, research can be redirected toward policies that attempt to cushion the effects of the successive crises.

The following section presents the new similarity index and an application of this index is given thereafter, by calculating the similarity between a set of lines in the economy area. The last section is devoted to conclusions and future work, with some ideas about various fields where it seems possible to develop this index.

Similarities Between Events

Similarity must be understood, initially, as a measure between two elements in a set X, which provides a numerical value to quantify how analogous they are. An important kind of similarity measures are dot products. In order to be able to use a dot product as a similarity measure in X, this domain must be embedded into some dot product space H. To this end, a map is used φ: X→H. Thus, a particular similarity measure in X, called Mercer kernel, is defined as

$$k: X\times X\rightarrow R$$
$$(x,x')\rightarrow k(x,x')=\langle\varphi(x),\varphi(x')\rangle_H$$

where dot product in H is denoted by $\langle\bullet,\bullet\rangle$. The idea of a kernel generalizes the dot product in the space X and provides a descriptive language used by the learning machine to see the data. Kernels offer a solution for projecting a dataset in a large feature space, which also increases the ability to generalize the various training algorithms. On the other hand, kernels can be interpreted as a similarity measure and this is the way they will be used in this paper. Hence, a kernel on a probabilistic space is to be considered.

Let us consider two events A and B on a (Ω,A,P) probabilistic space. The similarity between them, denoted by k(A,B) is defined González *et al*. (2005) as

$$k(A,B)= P(A,B)-P(A)P(B)$$

This similarity can take positive and negative values, and has interesting properties, as can be seen in González *et al*. (2005). Nevertheless, it is worth noting that this similarity has a great drawback, since $k(A,B) \leq k(A,A)= P(A)(1-P(A))$ and, therefore, if P(A) is small, then k(A,B) will be small for any B. Thus, for example, if P(A)= 0.01 and P(B)= 0.5, then k(A,A)= 0.009 and k(B,B)= 0.25. That is, B is approximately 28 times more similar to B than A to A and this result is not logical. To avoid this drawback, a new index is considered:

$$k^*(A,B)= \frac{k(A,B)}{\sqrt{k(A,A)k(B,B)}}$$

The most important property of k* is that it is the correlation between two random variables, $k^*(A,B)= corr(I_A,I_B)$, where I_A and I_B are the indicator functions of A and B, respectively, and corr stands for correlation. This result is straightforward to prove, since $k(A,B)= cov(I_A,I_B)$, where cov is the covariance (González *et al*., 2005). Therefore $-1\leq k^*(A,B)\leq 1$ and bounds are attained with $B=\bar{A}$ and B=A. Hence, $-1=k^*(A,\bar{A})\leq k^*(A,B)\leq k^*(A,B)\leq 1$ for any A and B; that is, the most similar event to A is the same A, and the most dissimilar event to A is the complementary event of A ($\bar{A}$). Therefore, k* does not have the drawback of k and this is also normalized. On the other hand, it is well-known that the correlation is a dot product (Schölkopf and Smola, 2002) and therefore, k* is a kernel.

Similarities Among Several Research Lines in Economics

The research lines shown in Table I have been obtained from the SSCI (Social Science Citation index) database in the ISI (Institute for Scientific Information) web of Knowledge (Thompson, 1945-2008). The years from 1990 to 2003, both included, have been selected for this study.

TABLE I
THE THIRTEEN RESEARCH LINES RELATED TO THE AREA OF ECONOMICS

Notation	Research Lines	Notation	Research Lines
A_1	Business	A_8	Transportation
A_2	Economics	A_9	Urban Studies
A_3	Environmental Studies	A_{10}	Social Sciences
A_4	Family Studies	A_{11}	Labor
A_5	Management	A_{12}	History
A_6	Planning	A_{13}	Finance
A_7	Development		

TABLE II
DOUBLE ENTRANCE TABLE WITH n_{IJ} ALONG THE 1990-2003 PERIOD

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	25545	754	766	623	4519	976	3708	153	520	51	1036	726	475
A_2		17741	973	321	1234	282	1745	107	341	179	629	725	302
A_3			30699	1660	3925	1473	5437	284	1442	98	356	1483	111
A_4				60360	2961	3026	6603	142	2006	57	1631	6027	120
A_5					66333	3960	10203	434	1550	120	1386	1946	530
A_6						24366	5081	542	2260	44	252	739	133
A_7							114487	544	4711	343	1646	5306	746
A_8								4273	601	4	129	85	44
A_9									26337	50	832	1238	202
A_{10}										2783	29	241	11
A_{11}											21926	873	156
A_{12}												61234	129
A_{13}													5687

Preliminary study

A total of N= 372805 different papers have been found after a thorough search, each of them related to at least one of the 13 research lines considered. Table II shows the values obtained for n_{ij} , the number of papers appearing simultaneously in lines Ai and Aj for the period considered, 1990-2003. Also, the papers published yearly for each of the 13 research lines from 1990 to 2003 have been depicted in Figure 1.

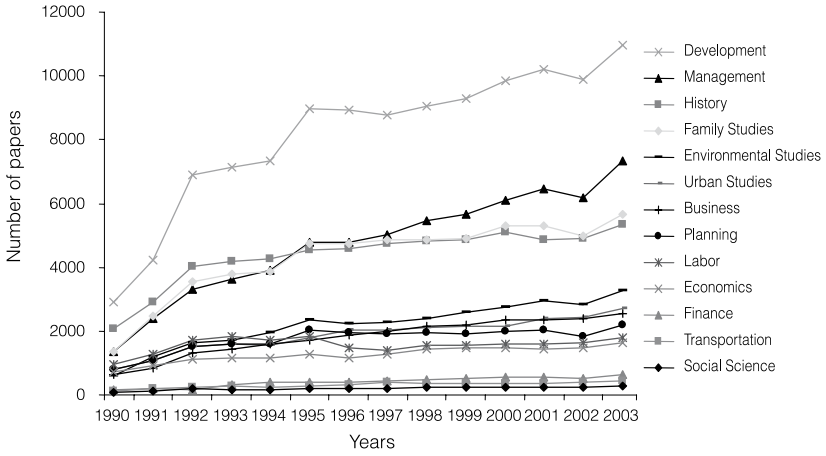

Figure 1. Graphical representation of yearly published papers for each of the 13 research lines from 1990 to 2003.

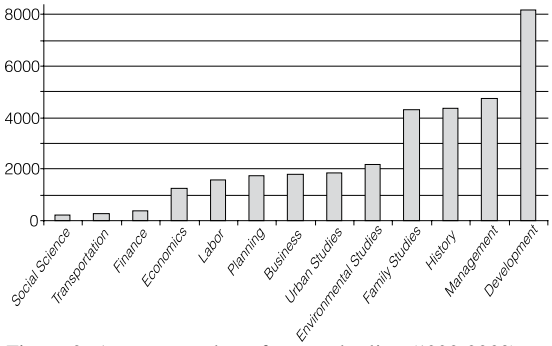

Figure 2. Average number of papers by line (1990-2003).

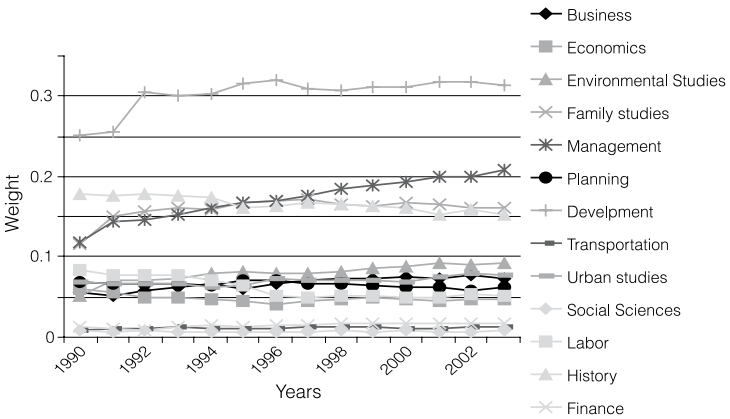

Figure. 3. Weight of every topic along the 1990-2003 term.

It can be observed from Table II and Figure 1 that Development (A_7) is the line with the most published papers, with a total of 114487, whereas Social Science (A_{10}), with 2783, is the line with the fewest published papers. Furthermore, it can be seen that Development has always been the most published line, year after year, and that Social Science is the fewest published line year after year except in 1992 when it was Finance.

Figure 1 also shows that the order of research lines change in time when they are sorted according to the amount of published papers. The published paper's average for each research line is shown in Figure 2, where they are ordered from the smallest to the greatest in the period of study. It can be seen that Social Sciences is the line with fewest papers (on average) with a value of 198.77, while the line with most papers is Development, with an average of 8177.64.

Another aspect to emphasize is the different weights for each research line according to their amount of published papers; that is, the proportion of each line with respect to the total number of published papers expressed over 1. These weights are shown in Figure 3 for every year. Hence, for example, Development takes in 1990 a value of 0.2508, resulting from the fact that this year Development has 2909 out of 11601 papers. Figure 3 shows clearly that some line weights change through the years, such as, for instance, happens with Family Studies and Management. Family Studies is above Management from 1990 to 1994, both are very close between 1995 and 1997 and from 1998 Management overtakes the other line.

Nevertheless, in this preliminary study, the hypothetical similarities among the research lines do not appear clearly, in the sense that it is not known how are the relations between lines. Thus, a study of similarities is carried out.

Study of the similarities

It is worth noting that Table II reflects the relations between A_i and A_j in absolute values (n_{ij} for $i \neq j$), but it does not show if there is any similarity or dissimilarity between them. Thus, the relationship between the research lines shown in Table I is studied according to the similarity measure k*.

To link the data from Table II to the construction of similarities, keeping in mind that the number of data available is sufficiently large, a frequency-based interpretation

TABLE III
K* NORMALIZED SIMILARITIES AMONG THE RESEARCH LINES ALONG THE 1990-2003 PERIOD

	P(Ai)	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	0.0685	1	-0.023	-0.052	-0.101	-0.001	-0.030	-0.095	-0.014	-0.053	-0.017	-0.021	-0.099	**0.0074**
A_2	0.0476		1	-0.022	-0.087	-0.063	-0.045	-0.101	-0.011	-0.045	**0.0068**	-0.022	-0.074	**0.0032**
A_3	0.0823			1	-0.088	-0.039	-0.021	-0.084	-0.006	-0.028	-0.015	-0.060	-0.094	-0.028
A_4	0.1619				1	-0.148	-0.027	-0.188	-0.037	-0.064	-0.033	-0.059	-0.076	-0.048
A_5	0.1779					1	-0.011	-0.155	-0.022	-0.086	-0.031	-0.075	-0.169	-0.028
A_6	0.0654						1	-0.057	**0.0268**	**0.0228**	-0.017	-0.055	-0.096	-0.021
A_7	0.3071							1	-0.042	-0.076	-0.035	-0.126	-0.212	-0.047
A_8	0.0115								1	**0.0294**	-0.008	-0.013	-0.042	-0.004
A_9	0.0703									1	-0.018	-0.032	-0.087	-0.017
A_{10}	0.0075										1	-0.018	-0.018	**0.0028**
A_{11}	0.0588											1	-0.084	-0.016
A_{12}	0.1643												1	-0.048
A_{13}	0.0153													1

of probability is considered. Hence, assuming that

$$P(A_i \cap A_j) = \frac{n_{ij}}{N}$$

where $N = \sum_{i,j} n_{ij}$, k* is used as in Eq. 1 to calculate the similarity between lines. These calculations are shown in Table III, and their similarities are depicted in Figure 4, where the positions for each item is shown with respect to the others using the values when k*(Ai)[Aj] goes from i= 1 to 13 (see Table I). Hence, it can be seen that the similarities between the different lines are negative, except in some specific cases. This was expected because if two lines are very similar it means that both lines share many papers and they should be considered as the same line. The greatest value of the similarity index in the whole period is $k^*(A_8,A_9)=0.02942$, linking Transportation and Urban Studies. The greatest similarity value for each year is depicted in Figure 5, where it can be seen that the greatest similarity value is given for the similarity between Planning and Transportation in 2003 ($k^*(A_6,A_8)=0.04893$).

It should be pointed out that the graphical scales in Figure 4 are different. In the y-axis the maximum depicted for each line is 0.08, 0.05, 0.08, 0.15, 0.20, 0.08, 0.25, 0.012, 0.08, 0.008, 0.06, 0.15 and 0.02, and it must be taken into account that each of these representations refers to one particular line.

Looking again at Table III and Figure 4, the following conclusions can be drawn about the similarity between each research line and its links to the others:

A_1 has positive similarity with A_{13} (0.00739) and is dissimilar to the others. The most dissimilar items to A_1 are A_4 and A_{12}, with analogous dissimilarity levels. The remaining dissimilarities are quite small.

A_2 has positive similarity with A_{13} and A_{10} and is dissimilar to the rest. The greatest similarity is related to A_7.

A_3 has dissimilarities with all the other lines. The most dissimilar is A_{12}.

A_4 has dissimilarities with all the other lines. The most dissimilar is A_7.

A_5 has dissimilarities with all the other lines. The most dissimilar is A_{12}.

A_6 has positive similarity with A_8 (0.02679) and A_9 (0.02281). The other lines are all dissimilar and the most dissimilar is A_{12}.

A_7 shows negative similarity to the other lines, being A_{12} the most dissimilar.

A_8 has positive similarity with A_6 (0.02679). It is dissimilar to the other lines, and A_{12} is the most dissimilar line.

A_9 has positive similarity with A_6 (0.02281) and A_8 (0.02679). A_{12} is the most dissimilar.

A_{10} has positive similarity with A_2 (0.00682) and also with A_{13} (0.00280) and A_7 is its most similar line.

A_{11} has negative similarity with each of the other lines and A_7 is the most dissimilar line.

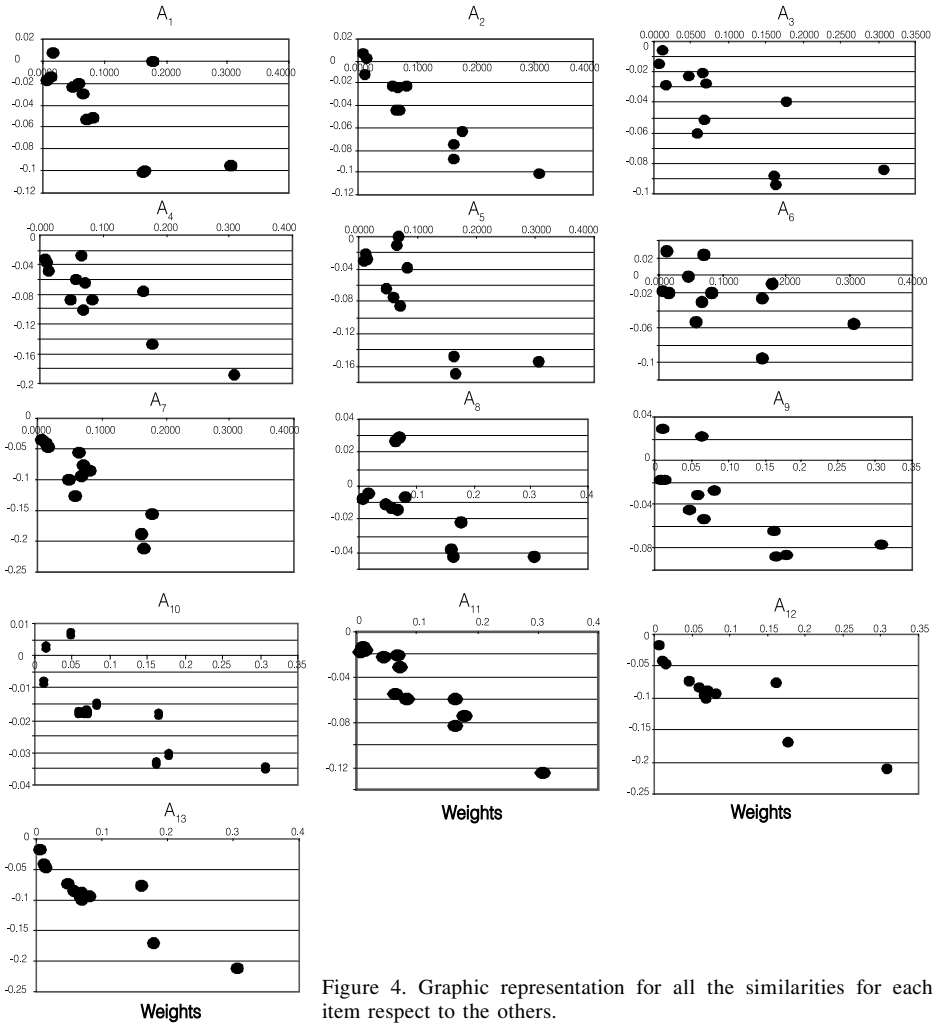

Figure 4. Graphic representation for all the similarities for each item respect to the others.

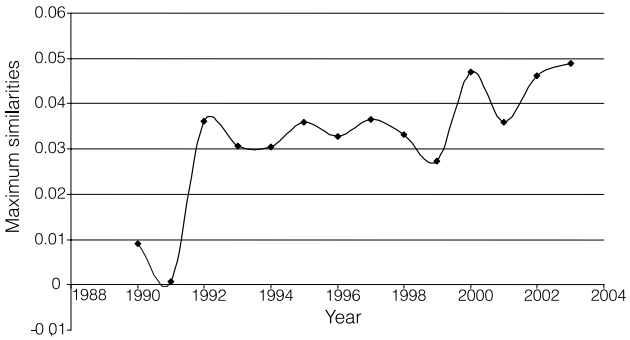

Figure 5. Maximum similarities representation along the 1990-2003 period.

TABLE IV
ORDER OF EACH RESEARCH LINE RELATED TO THE OTHERS ACCORDING TO THE SIMILARITY

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
1	A_4	A_7	A_{12}	A_7	A_{12}	A_{12}	A_{12}	A_7	A_{12}	A_7	A_7	A_7	A_4
2	A_{12}	A_4	A_{14}	A_5	A_7	A_7	A_4	A_{12}	A_5	A_4	A_{12}	A_5	A_{12}
3	A_7	A_{12}	A_7	A_1	A_4	A_{11}	A_5	A_4	A_7	A_5	A_5	A_1	A_7
4	A_9	A_5	A_{11}	A_3	A_9	A_2	A_{11}	A_5	A_4	A_{12}	A_3	A_6	A_3
5	A_3	A_9	A_1	A_2	A_{11}	A_1	A_2	A_1	A_1	A_{11}	A_4	A_3	A_5
6	A_6	A_6	A_5	A_{12}	A_2	A_4	A_1	A_{11}	A_2	A_9	A_6	A_9	A_6
7	A_2	A_1	A_{13}	A_9	A_3	A_{13}	A_3	A_2	A_{11}	A_6	A_9	A_{11}	A_9
8	A_{11}	A_3	A_9	A_{11}	A_{10}	A_3	A_9	A_{10}	A_3	A_1	A_2	A_4	A_{11}
9	A_{10}	A_{11}	A_2	A_{13}	A_{13}	A_{10}	A_6	A_3	A_{10}	A_3	A_1	A_2	A_8
10	A_8	A_8	A_6	A_8	A_8	A_5	A_{13}	A_{13}	A_{13}	A_8	A_{10}	A_{13}	A_{10}
11	A_5	A_{13}	A_{10}	A_{10}	A_6	A_9	A_8	A_6	A_6	A_{13}	A_{13}	A_8	A_2
12	A_{13}	A_{10}	A_8	A_6	A_1	A_8	A_{10}	A_9	A_8	A_2	A_8	A_{10}	A_1

A_{12} has negative similarity with the other lines, as in the previous case, being A_7 the most dissimilar.

A_{13} by symmetry with A_1, A_2 and A_{10} has positive similarity with all them and is dissimilar to the others. A_4 is the most similar line.

As it can be seen, A_7 (Development) is the most dissimilar one to the other research lines. This point shall be brought up later, after explaining the construction of Table IV. It should be remarked that the main deficiency of Figure 4 is the use of so many figures as studied items. It would be much more practical to have only one representation to gather all the previous graphic information. To fulfill this aim, the following graphical construction is proposed:

Firstly, the events $\{A_1,A_2, ...,A_{13}\}$ are ordered by their probabilities; that is, $0 \leq P(A_1) \leq P(A_2) \leq ... \leq P(A_{13})$ and these probabilities are represented on the x-axis. The similarities are represented on the y-axis in the following way: first, for the event A_1 the set of similarities $\{-k(A_1,A_1), k(A_1,A_2),..., k(A_1,A_{13}), k(A_1,A_1)\}$ is represented on the abscissa $P(A_1)$, and then for the event A_2 the set of similarities $\{-k(A_2,A_2), k(A_2,A_3), ..., k(A_2,A_{13}), k(A_2,A_2)\}$ is represented on the abscissa $P(A_2)$, and so on up to the event A_{13}. The graphical representation, following this procedure applied to the 13 lines of research related to Economy may be seen in Figure 6. To determine which lines are most similar (or dissimilar) to the others, Table IV has been built, where similarity is represented in columns in increasing order; then, for instance, in the A_1 column, A_4 is the most dissimilar one to it, followed by A_{12} and so on up to A_{13} where the maximum similarity is found (the first picture in Figure 4 shows this situation). Nevertheless, the order obtained does not match with the absolute values of Table II and, for instance, line A_{10} is the first in the A_1 column in Table II in relation to the number n_{ij}, although A_{10} is in the 9th position in that column in Table IV, and A_4 is situated in the 5th position in Table II and in the first place in Table IV. It is remarkable that A_{12} is in the 11th place in Table II and the second in Table IV. This fact can be explained since the considered values in Table IV are relative but not absolute values as in Table II. Hence, Table IV gives a more suitable order relation according to the affinity among the topics.

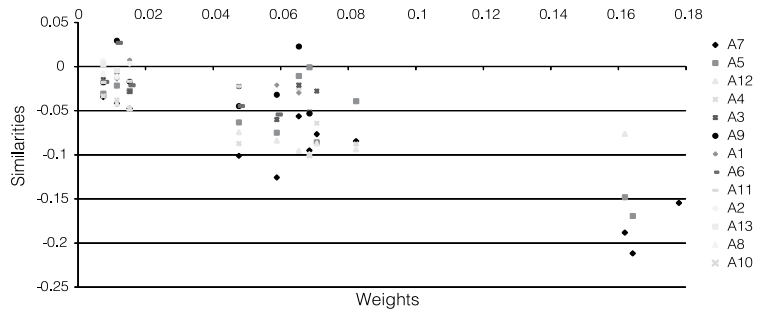

Figure 6. Graphical representation of all similarities.

TABLE V
DISSIMILARITY OF EACH RESEARCH LINE, SUM, RANK, AND AVERAGE RANK OF EACH RESEARCH LINE WITH RESPECT TO THE OTHERS

	1	2	3	4	5	6	7	8	9	10	11	12	13	Sum	Average	Av. rank/year
A_7	6	2	4	-	-	-	-	-	-	-	-	-	1	35	2.7	2.6
A_{12}	5	4	1	1	-	1	-	-	-	-	-	-	1	39	3.0	3.2
A_4	2	4	2	1	1	1	-	1	-	-	-	-	1	52	4.0	4.1
A_5	-	3	3	2	1	1	-	-	-	1	1	-	1	68	5.2	5.2
A_{11}	-	-	1	2	2	1	2	3	1	-	-	-	1	87	6.7	6.5
A_3	-	-	-	3	2	-	2	3	2	-	-	-	1	91	7.0	7.0
A_1	-	-	2	-	4	1	1	1	1	-	-	2	1	93	7.2	7.3
A_9	-	-	-	2	1	2	3	2	-	-	1	1	1	98	7.5	7.6
A_2	-	-	-	1	2	2	2	1	2	-	1	1	1	102	7.8	7.8
A_6	-	-	-	1	-	4	1	-	1	1	3	1	1	112	8.6	8.6
A_{13}	-	-	-	-	-	-	2	-	2	4	3	1	1	130	10.0	10.0
A_{10}	-	-	-	-	-	-	-	2	3	2	2	3	1	134	10.3	10.3
A_8	-	-	-	-	-	-	-	-	1	5	2	4	1	142	10.9	10.7

Table IV has some aspects that are shown in Table V. On the first row of Table IV it can be seen, that the A_7 appears six times, A_{12} five times, and A_4 twice. These numbers appear in this order in the second column of Table V, indicating that they are in the first dissimilarity position. So, this is the order of the topics in the first column. Analogously, in the second row of Table IV A_7 appears twice, A_{12} four times, A_4 four times, and A_5 three times, and these appear in the third column of Table V. That is, Table V indicates the dissimilarity of each topic in relation with the others, not taking into account the similarity values, but the order. To be able to make comparisons a weighted sum is carried out, where the weights are the order positions from the first row, and the results are written in the antepenultimate column. In this way, if for instance A_{12} is considered, the result $5\times1+4\times2+1\times3+1\times4+1\times6+1\times13=39$ is obtained and placed in the antepenultimate column of Table V. After dividing into 13, the weight of the topic is found, referred to its dissimilarity with respect to the other ones.

TABLE VI
MAXIMUM SIMILARITIES ALONG THE 1990-2003 PERIOD

Year	Similarity	Pos. sim. num.	Lines
1990	0.009015	1	A_8-A_9
1991	0.000782	1	A_6-A_8
1992	0.036096	6	A_1-A_{13}
1993	0.030621	7	A_8-A_9
1994	0.030535	5	A_8-A_9
1995	0.035925	5	A_8-A_9
1996	0.032662	6	A_6-A_8
1997	0.036425	6	A_8-A_9
1998	0.033272	5	A_8-A_9
1999	0.027378	6	A_6-A_9
2000	0.046976	6	A_6-A_8
2001	0.035831	7	A_6-A_8
2002	0.046076	5	A_6-A_9
2003	0.048929	5	A_6-A_8

In the case of A_{11}, the weighted sum is 87. Arranging the table according to the last column Table V is constructed, where the topics are sorted from lower to higher similarity. From this Table it can be seen that the item more related to the others is A_8, followed by A_{10} and finishing with A_{12} and A_7 , respectively. This Table shows that the strongest subjects are A_7, A_{12}, A_4 and A_5, in the sense of being the most published and because they are in the first rows of Table V, thus having more dissimilarity with the others. This result indicates that they share little with the other lines and seem to be self-sufficient. Table V has a complementary column where average rank by year is calculated.

It was stated before that the lines with highest similarity for the whole period were A_8 and A_9 , with a value of 0.02942, and it can be questioned whether this situation holds for every year. Figure 5 provides a representation of the maximum similarities in the 1990-2003 period. These similarities have followed an increasing trend, although the lines are not always the same. Table VI shows the lines for which the similarities are maximal each year. Line A_8 appears in 11 years, A_9 appears in 8 years, A_6 in seven years and A_{13} and A_1 once each. Also, the pair formed by A_8 and A_9 appears six times, A_6 with A_8 five times, A_6 with A_9 twice, and, finally, A_1 with A_{13} once. This means that A_8 and A_9 are really the lines with the highest similarity between them. Figure 7 shows the similarities between A_8 and A_9 in the 1990-2003 period. As it is possible to see in this Figure, in 1991 the similarity is negative, changing to positive in 1992. The trend is ascendant, with slight punctual variations in some years.

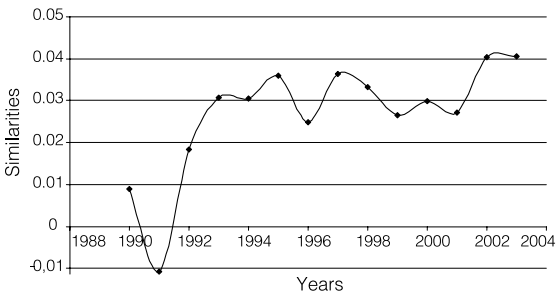

Figure 7. Similarities between A_8 and A_9 along the 1990-2003 period.

Conclusions and Future Work

A new index of similarity is presented, improving the one presented by González *et al*. (2005). In addition, it has been shown that the similarity between events can be used to detect that, although the output of papers in some research areas of Economy seems to be high, this is not the case in other areas in the same field. In the present paper it has been shown that lines such as Development (A_7), with 114487 published papers in the 1990-2003 period, do not have great similarity with the other lines. This is because the number of work coincidences with other areas is proportionally very small. This also holds for other important lines in relation to the amount of published papers, such as History (A_{12}), Management (A_5) and Family Studies (A_4). Throughout all this work it is seen that these lines are quite independent. It has been stated, in another way, that Social Science (A_{10}) is the line with fewest papers in Table II but, however, it holds a good relationship with the other lines, and thus its position in Table V is logical.

The different lines can also be sorted according to the dissimilarities (similarities), indicating which ones are most similar to the others.

Several lines of work are planned for the near future. One of them will consist on using the science databases as documental information to carry out a similar study into different subjects in areas of interest. Another work line will consist on making a dynamic study about the relative growth of each subtopic and the existing temporal similarities between topics and subtopics.

It is also envisaged to find out the similarities between the relevant groups of research in certain areas, sub-areas or research lines on the web. In this paper, it has been attempted to highlight some of the applications that can be developed with the mathematical method of González (2002) and González *et al*. (2005), briefly presented above. This gives an idea of the power of this similarity index, whose potential possibilities are being extended to real intervals and to time series, also allowing the study of similarities in cases such as economic series, radio or television audiences, etc.

ACKNOWLEDGMENTS

This work has been partially financed by the PAI-2009/00001200, PAI-2008/00000607, P06-TIC-02141 help, provided by the Junta de Andalucía, and by the Science and Technology Ministerial Department under project TIN2009-14378-C02-01.

REFERENCES

Almind T, Ingwersen P (1997) Informetric analyses on the world wide web: methodological approaches to webometrics. *J. Docum. 53*: 404-426.

Baños RR, Contreras F (1998) Como consultar eficazmente una base de datos bibliográfica. El método de las palabras asociadas. www.ugr.es/~fccortes/curriculum/toledo.html (In spanish).

Braam R, Moed H, Raan AV (1991) Mapping of science by combined cocitation and word analysis. ii: dynamical aspect. *Am. Soc. Inf. Sci. 42*: 252-266.

Burrell QL (2005) Measuring similarity of concentration between different informetric distributions: Two new approaches. *J. Am. Soc. Inf. Sci. Technol. 56*: 704-714.

Buter R, Noyons E (2002) Using bibliometric maps to visualise term distribution in scientific papers. In *Proc. 6th Int. Conf. on Information Visualisation*. pp. 697-702.

Callon M, Law J, Rip A (1986) (Eds.) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Macmillan. London, UK.

Callon M, Courtial J, Laville F (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics 22*: 155-205.

Clara N (2006) Generalized fuzzy similarity indexes. *Lecture Notes in Computer Science 393*1: 163-170.

Coulter N, Monarch I, Konda S (1998) Software engineering as seen through its research literature: A study in co-word analysis. *Am. Soc. Inf. Sci. 49*: 1206-1223.

Courtial J (1994) A co-word analysis of scientometrics. *Scientometrics 3*: 251-260.

Deus J (2001) *Escalamiento Multidimensional*. La Muralla. Madrid, España. 144 pp.

Egghe L, Rousseau R (1990) *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*. Elsevier. Amsterdam, Holland.

González L (2002) *Análisis Discriminante Utilizando Máquinas Núcleos de Vectores Soporte. Función Núcleo Similitud*. Thesis. Universidad de Sevilla. Spain.

González L, Velasco F, Gasca R (2005) A study of the similarities between topics. *Comput. Stat. 20*: 465-479.

Gonzalez-Abril L, Cuberos FJ, Velasco F, Ortega JA (2009a) Ameva: An autonomous discretization algorithm. *Expert Syst. Applic. 36*: 5327-5332.

Gonzalez-Abril L, Velasco F, Ortega JA, Cuberos FJ (2009b) A new approach to qualitative learning in time series. *Expert Syst. Applic. 36*: 9924-9927.

Grivel L, Francois C (1995) Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique. *Solaris 1995*: 81-112.

Kinnucan M, Nelson M, Allen B (1987) Statistical methods in information science research. *Annu. Rev. Inf. Sci. Tecnol. 22: 147-178*.

Klock H, Buhman J (1999) Data visualization by multidimensional scaling: A deterministic annealing approach. *Pattern Recogn. 33*: 651-669.

Kohonen T (1998) Self-organization of very large document collections: State of the art. *Proc. ICANN98*: 65-74.

Kohonen T, Kaski S, Lagus K, Salojarvi J, Honkela J, Paatero V, Saarela A (2000) Self-organization of a massive document collection. *IEEE Trans. Neural Net. 11*: 574-585.

Larson R (1996) Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace. http://sherlock.berkeley.edu/asis96/asis96.html.

Li Y, Olson DL, Qin Z (2007) Similarity measures between intuitionistic fuzzy (vague) sets: A comparative analysis. *Pattern Recogn. Lett. 28*: 278-285.

Lin X, Marchionini G (1991) A self-organizing semantic map for information retrieval. In *Proc 14 ACM/SIGIR Conf. Research and Development in Information Retrieval*. Proc. ACM SIGIR'91, Chicago, published as a special issue of SIGIR FORUM, ACM Press, pp 262-269.

Mijac V, Ryder E (2009) Análisis bibliométrico de las poblaciones científicas sobre parasitosis en Venezuela (2002-2007). *Interciencia 34*: 140-146.

Noyons E, Buter R, Raan AV (2002) Bibliometric mapping as a science policy tool. In *Proc. 6th Int. Conf. on Information Visualisation*. pp. 679-684.

de la Rosa F, Gasca R, González L, Velasco F (2005) Análisis de redes sociales mediante diagramas estratégicos y estructurales. *Redes 8*: 5-37.

Rousseau R (1997) *Citations: An Exploratory Study*. Thesis. KHBO-Industrial Sciences and Techonology Zeedijk. Oostende, Belgium.

Schölkopf B, Smola AJ (2002) *Learning with Kernels*. MIT Press. Cambridge, MA, USA.

Srebro N (2007) How good is a kernel when used as a similarity measure? *Lecture Notes in Artificial Intelligence 4539*: 323-335.

Stentiford F (2007) Attention-based similarity. *Pattern Recogn. 40*: 771-783.

Thomson (1945-2008) *ISI Web of Knowledge*. www.isinet.com/journals/.

Vapnik V (1998) *Statistical Learning Theory*. Wiley.

Wolfram D (2000) Applications of informetrics to informetrics to information retrieval research. *Inf. Sci. 3*: 77-82.

Zhang C, Fu H (2006) Similarity measures on three kinds of fuzzy sets. *Pattern Recogn. Lett. 27*: 1307-1317.

UN ESTUDIO DE ÁREAS TEMÁTICAS EN ECONOMÍA CON UNA MEDIDA DE SIMILITUD BASADA EN LA TEORÍA DE LOS NÚCLEOS

Francisco Velasco, Luis González-Abril, Juan Antonio Ortega y Juan Antonio Álvarez

RESUMEN

Presentamos en este trabajo resultados interesantes acerca de las similitudes existentes entre trece tópicos dentro del ámbito de la Economía entre los años 1990 y 2003, así como las interrelaciones existentes entre ellas. Para ello, introducimos una medida de similitud intuitiva entre dos conjuntos, basada en la teoría de los núcleos, que nos permite dar una representación gráfica muy útil. Consideramos además varias tablas en las que se muestra la potencialidad futura del índice de similitud introducido.

UM ESTUDO DE ÁREAS TEMÁTICAS EM ECONOMIA COM UMA MEDIDA DE SIMILARIDADE BASEADA NA TEORIA DOS NÚCLEOS

Francisco Velasco, Luis González-Abril, Juan Antonio Ortega e Juan Antonio Álvarez

RESUMO

Apresentam-se resultados sobre as similaridade existentes entre treze tópicos dentro do âmbito da Economia entre os anos 1990 e 2003, assim como as interrelações existentes entre eles. Para isto, se introduz uma medida de similaridade intuitiva entre dois conjuntos, baseada na teoria dos núcleos, que permite dar uma representação gráfica útil. Se consideram além disso varias tabelas nas que se mostra a potencialidade do índice de similaridade introduzido.