A Web Map of the CSIC Research Centres: A comparative study of the Cosine and the Pearson's Correlation Coefficient in a Colink Analysis

José Luis Ortega Priego and Isidro Aguillo

jortega@cindoc.csic.es, isidro@cindoc.csic.es
Internet Lab, Centro de Información y Documentación Científica (CSIC),
Joaquín Costa, 22, 28002 Madrid (Spain)

Abstract
A colink study of the web sites of the different centres of the Spanish main public research body is made. The purpose is to find the relations existing among research areas of these centres as well as to study the similitudes and differences according to two measures: cosine and Pearson's correlation coefficient. A colink matrix is built from Yahoo! Search results. The main research areas identified in the CSIC are the Physics and Materials Sciences areas and the Agrobiology, Biomedicine and Food Technologies areas, proving a greater importance in the applied research than in the fundamental research. With regard to the results in the cosine and the correlation coefficient model there are slight differences between two measures as much in the MDS map as in the clustering dendrogram.

Introduction

Recently, it gave rise to a debate about the suitability of the Pearson's correlation coefficient and the cosine of Salton. Ahlgren, et al. (2003) showed that the Pearson's correlation coefficient is not a suitable measurement for ACA (Author Cocitation Analysis) since it is sensitive to the number of zeros and for new added variables. However, White (2003) argued that this inconsistence did not affect the final ACA result, the data clustering and graphical mapping. Bensman (2004) expressed that the arguments of the paper are theoretically in excess and these are not significative in a real-world example. Ahlgren, et al. proposed several alternative measures (cosine, chi-squared) while Leydesdorff (2004a) suggested the information theory as a suitable clustering methodology. However, different works have chosen the use of cosine with satisfactory results (Leydersdorff 2004b).

In order to test this, we chose the Consejo Superior de Investigaciones Científicas (CSIC) as a target for analysis. The Spanish Council for Scientific Research (CSIC) was created in 1939. It comprises 131 research centres of different disciplines, some of them joint with universities. The presence of this centres in the web environment was studied previously by Ortega (2003) and Aguillo and Granadino (2004).

Objectives

The main objective was to display graphically the web sites of the CSIC Research Centres through a colink map in order to identify the main subject areas of this entity and the way in which these centres are interelated in a web environment. For this, we apply two different measures: the correlation coefficient and the cosine.

Methodology

The web sites of the CSIC research centres and laboratories have been selected for this analysis, and Yahoo! Search has been used to extract the colink raw matrix. 11 web sites have been removed since their URLs do not have a specific institutional subdomain, because the "linkdomain" operator of Yahoo! Search does not work with pages inside one domain. Sites without a research activity have been eliminated and for the centres that have merged, we have added up the colinks of both centres. The same for centres with several domains. Also, the web sites without any inlink have been removed. Finally, 111 web sites were used in this study.

The query syntax is following: "*+linkdomain:{domain} +linkdomain:{domain} –site:csic.es*" and has been used in a in-house developed script, in this way we can obtain the times that two web sites are colinked by other pages. Many researches have detected bias and unstability in the search engines results (Rousseau, 1999; Bar-Ilan, 2002; Thelwall, 2001; Vaughan & Thelwall, 2003), although the search engines keep being used as a suitable tool for colink studies due to their coverage and their improvement.

Proceedings of ISSI 2005, pages 197–204.
Edited by P. Ingwersen and B. Larsen.

Once the raw matrix was obtained, two different measures were applied to calculate the similarity: the Salton's cosine (Salton & McGill, 1983) and the Pearson's correlation Coefficient. The first can be defined as the cosine of angle between two vectors X and Y. This measure is not sensitive to the number of zeros as the cosine is not based in the mean of the distribution. The formula is:

$$\cos(A_1, A_2) = \frac{\sum_{i=1}^{N} CC_{1i} * CC_{2i}}{\sqrt{\left(\sum_{i=1}^{N} CC_{1i}^2\right)\left(\sum_{i=1}^{N} CC_{2i}^2\right)}}$$

where A_1 and A_2 are two web sites, and CC are the times that these web sites are colinked.

The Pearson's correlation coefficient (-r) describes the strength and direction of a linear relationship between two variables X and Y and can be defined as:

$$r = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2 \sum\left(Y - \overline{Y}\right)^2}}$$

where $\overline{X}$ and $\overline{Y}$ are the means of the two variables.

Finally, both results have been grouped by Agglomerative Hierarchical Clustering (AHC) and mapping through (MDS) Multidimensinal Scaling. The dendrogram has been created with the software XLStat 7.1 and the MDS maps have been elaborated with the PROXCAL module of SPSS 12.0.

Results

The diagonal values in the colink matrix (total external inlinks to individual sites) have produced a great distortion in the results, increasing significantly the MDS stress value and forming groups with little differences between themselves. Thus, we have considered the diagonal data with null values.

In the first MDS map (Figure 1) we can see that correlation coefficient spreads more the results and create more uniform groups. On the contrary, there is a high concentration of points in the centre of the MDS map calculated with cosine that make difficult its graphical grouping (Figure 2). In both models there is a high Stress value in the MDS map being higher in the correlation coefficient (φ=0.256) response than in the cosine (φ=0.233) one, but without significant differences. However, the correlation coefficient calculated 24 iterations until it would achieve a valid Stress value, while the cosine needed more than 50 iterations.

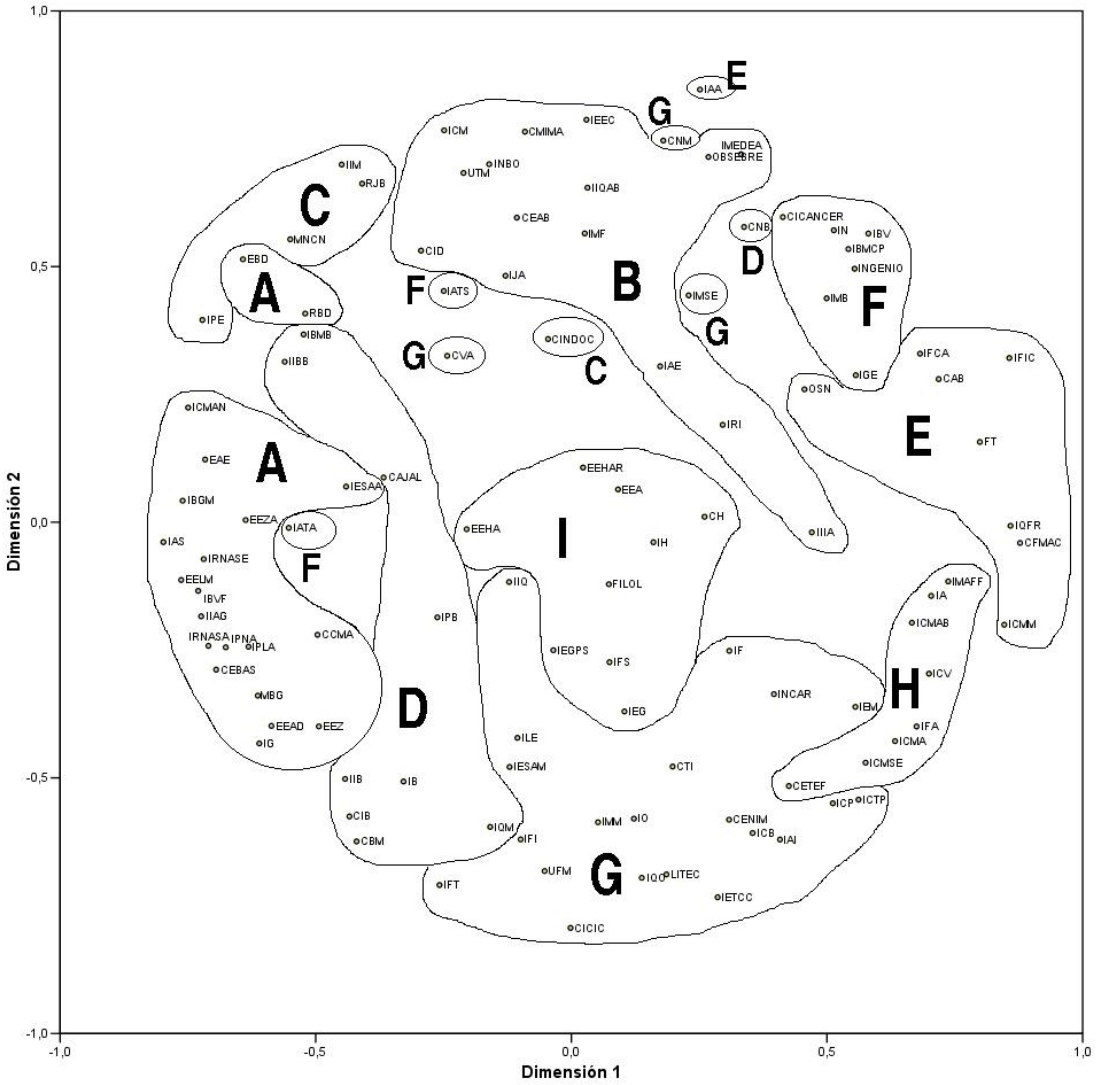

Figure 1. MDS map with correlation coefficient (φ=0.256).

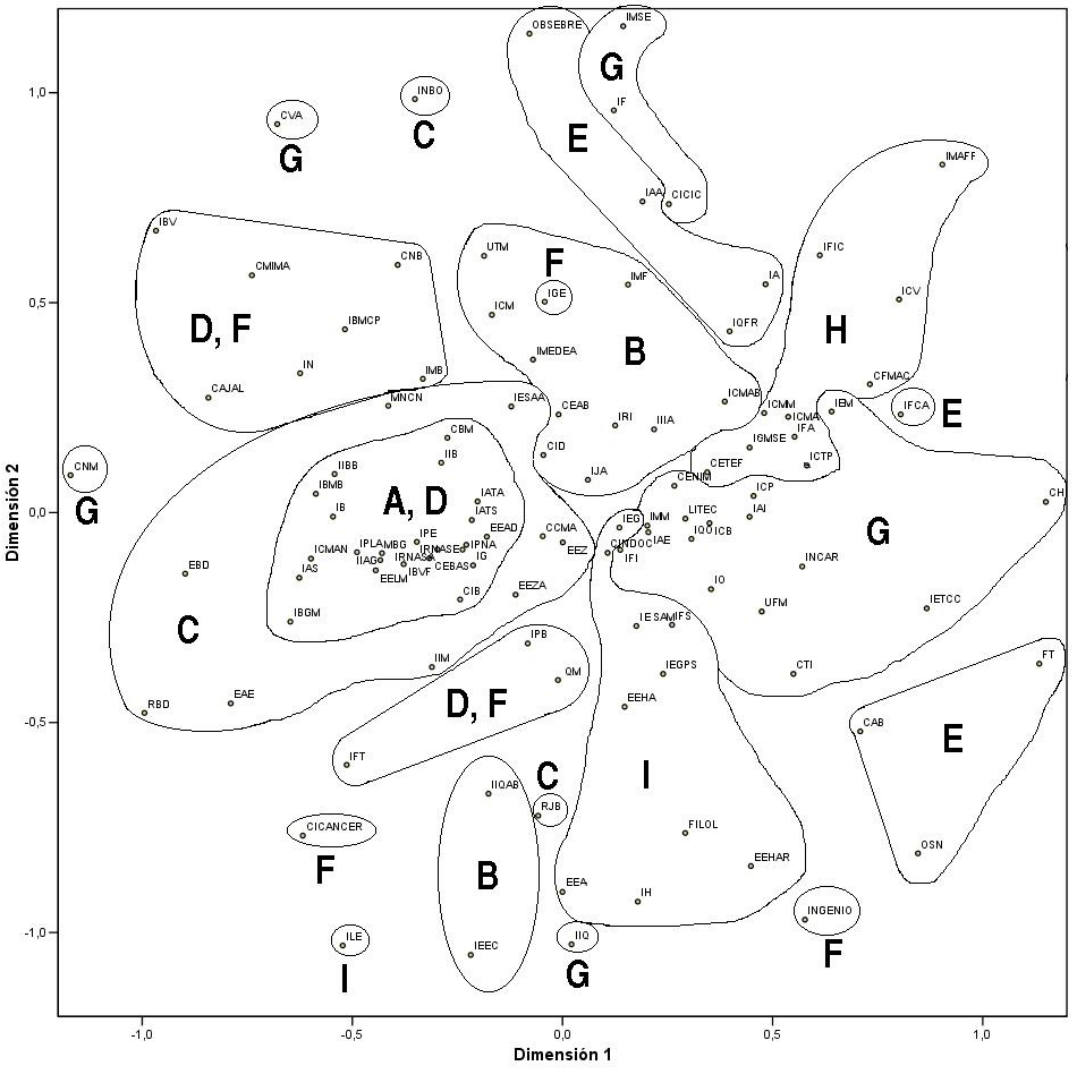

Figure 2. MDS map with cosine (φ=0.233).

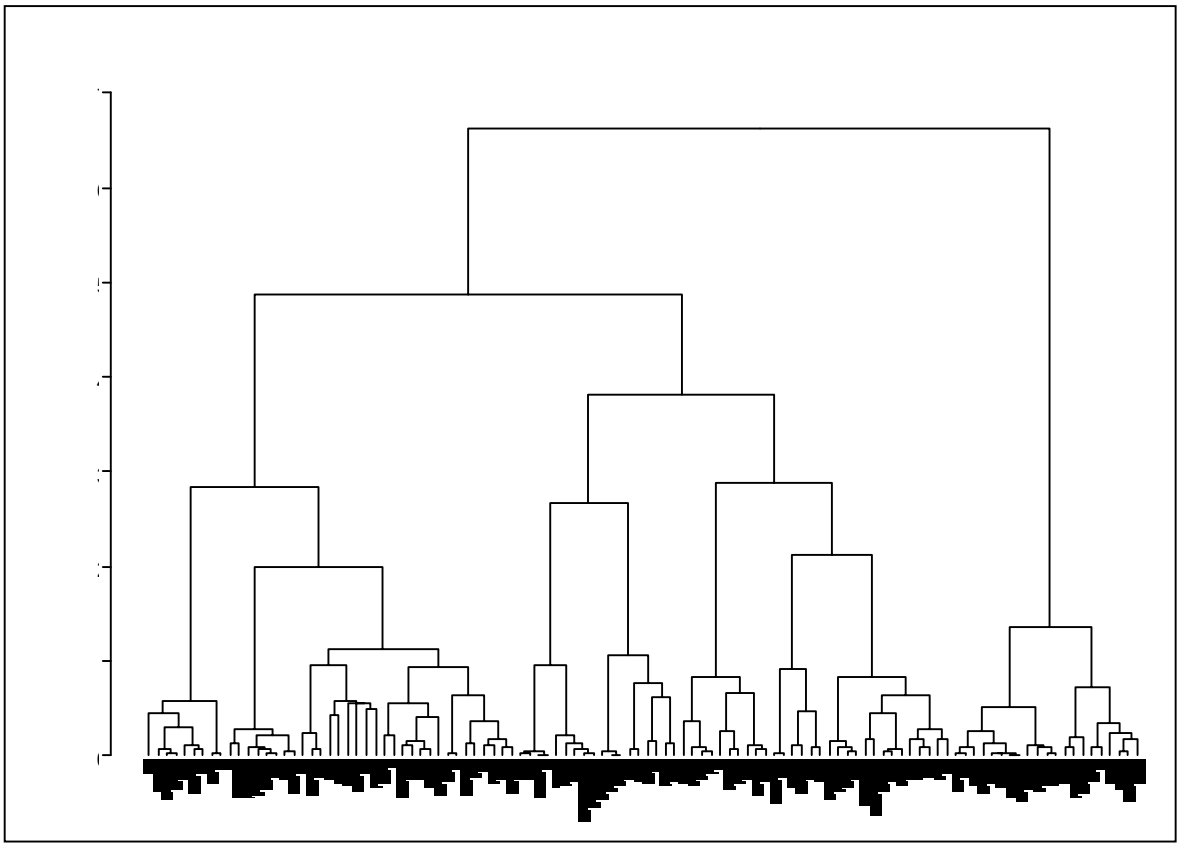

Figure 3. CSIC Dendrogram from correlation coefficient.

Table 1. CSIC Clusters from correlation coefficient.

Clusters	Research Centres	Disciplines
A	EELM, IAS, IBVF, CCMA, IIAG, IPNA, CEBAS, IRNASA, IBGM, EEAD, IPLA, MBG, EBD, ICMAN, IESAA, EEZ, IG, EAE, EEZA, IRNASE, RBD	Agrobiology, Plants Biology, Animals Biology
B	CMIMA, IIQAB, IEEC, IMF, IMEDEA, OBSEBRE, CEAB, ICM, UTM, IIIA, IAE, IJA, CID, IRI	Earth Sciences, Environmental Sciences, Catalonian
C	CINDOC, IIM, INBO, MNCN, IPE, RJB	Ecology, Botanics, Natural Sciences
D	CNB, CBM, CIB, IIB, IB, IBMB, IIBB, IPB, CAJAL, IQM	Biology, Biomedicine, Molecular Biology
E	CFMAC, ICMM, IQFR, IAA, IFIC, CAB, FT, IFCA, OSN	Physics, Astrophysics
F	IBV, IGE, IBMCP, IN, IATA, IATS, IMB, CICANCER, INGENIO	Biology, Medicine, Agrochemistry
G	IAI, IETCC, IFT, CTI, CVA, CNM, CICIC, IIQ, IMSE, UFM, IF, CENIM, ICB, ICP, ICTP, INCAR, IEM, IO, IESAM, ILE, IFI, IMM, IQO, LITEC	Engineering, Electronics
H	IA, ICMAB, CETEF, ICMSE, ICMA, IFA, ICV, IMAFF	Physics, Mathematics, Materials Sciences
I	CH, EEHA, EEHAR, FILOL, IEGPS, IFS, EEA, IH	Humanities

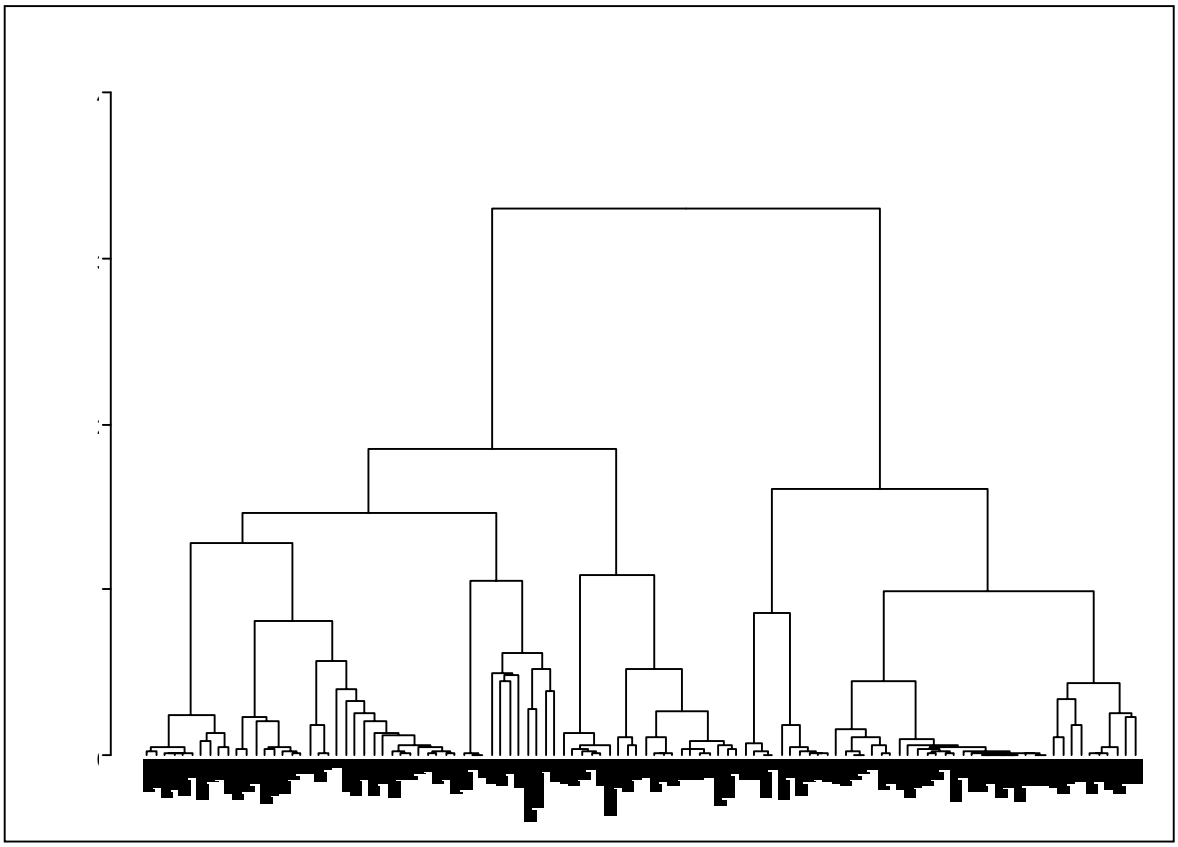

Figure 4. CSIC Dendrogram from cosine.

Table 2. CSIC Clusters from cosine.

Clusters	Research Centres	Disciplines
A, D	CNB, CBM, CIB, IIB, IB, IBMB, IIBB, IBGM, ICMAN, IATA, IATS, IPE, IG, IRNASE, IAS, EEAD, EELM, IIAG, CEBAS, IPNA, IRNASA, IBVF, IPLA, MBG	Biomedicine, Agrobiology
B	IEEC, IIQAB, IMF, IIM, CEAB, ICM, UTM, CID, IIIA, IAE, IJA, IMEDEA, ICMAB, IRI	Earth Sciences, Environmental Sciences, Catalonian
C	INBO, MNCN, EBD, RJB, CCMA, EEZ, EEZA, IESAA, EAE, RBD	Ecology, Botanics, Natural Sciences
E	CAB, IFCA, IAA, IA, IGFR, OBSEBRE	Astrophysics
D, F	IBV, IGE, IBMCP, IN, CMIMA, IFT, CAJAL, IMB, IPB, IQM	Biomedicine, Agrobiology
G	CH, IEM, IO, IF, CICIC, IETCC, CTI, INCAR, UFM, CENIM, ICB, ICP, IAI, IFI, IMM, IQO, LITEC	Engineering, Electronics
H	ICMM, ICTP, CETEF, ICMA, ICMSE, IFA, CFMAC, IFIC, ICV, IMAFF	Physics, Mathematics, Materials Sciences
I	EEHAR, EEHA, EEA, CINDOC, IESAM, IEGPS, IEG, IFS, FILOL, IH, ILE	Humanities

Figures 3 and 4 show the dendrograms built from the correlation coefficient and the cosine. The Ward's Method has been used for grouping the research centres web sites. The results in the MDS and in the dendrograms differ slightly. In Table 1 and 2 we can see the resultant cluster groups, this groups were labeled with the same capital letter in both tables.

Finally, the influence of a geographical variable can be appreciated in the Group B (Thelwall, 2002). All the centres of this group are located in Catalonian-speaking regions, similar tendencies were appreciated in the Group A with respect to Andalusia (South Spain).

Conclusions and Discussions

The results provided by of the cosine and the correlation coefficient are different (MDS maps and dendrograms). The cosine favors the similarities over the dissimilarities in the MDS display. In the cosine MDS map (Fig. 2) there is a great concentration of points in the centre whereas the points in periphery are more dispersed. However, the correlation coeffcient model (Fig. 1) presents less distortion between similarities and dissimilarities. Perhaps, this could be caused by the different ranges of the two measures, [0,1] in the cosine and [-1,1] in the correlation coefficient, and that this could affect the points disposition in the MDS map (Jones & Furnas, 1987). On the contrary, differences in the groups are either seen in clustering process.

White (2003) and Leydesdorff and Zaal (1988) did not see significant differences between the results of the two measures, but they used smaller samples (45 items for Leydesdorff & Zall and 24 items for White). In our case, we have used more than 100 items, so the size of the data could certainly have an influence on the results. In this form, the larger the size of the items of study, the greater could be the differences in the results.

Moreover, we used webometric data for this study, whereas in previous studies bibliometric data have been used. Several factors could affect the web data, such as the search engine stability, the presence of directories or lists, external variables (i.e. geography). For this reason, the meaning of the results could not be strictly considered. We would like to encourage forthcoming studies which can validate these measures according to the size of the analyzed sample and the type of data.

From our point of view and with certain caution, the results obtained from the correlation coefficient are better than the results contributed by the cosine, because the groups in the dendrogram are more solid, and in the MDS the layout is visually more clear and easy.

The areas identified match the findings by Aguillo and Granadino (2004). There are some areas that grouped most of the research centres. On one side, the Physics and Materials Sciences and on the other side Agrobiology, Biomedicine and Food Technologies. Thus, this could indicate that the CSIC tends to develop towards applied research and not to basic research. However, the little and limited presence of the Human and Social Sciences (there is a significant absence of areas such as Law, Politics and Psychology) indicates a great weight of the Natural and Technological Sciences. The presence of technological sciences could be due to the bigger web use by technicians rather than by other researchers (Aguillo, 2004). It also appreciates that there are multidisciplinary areas such as Chemical and Physical Technologies in which centres appear in several groups, indicating that some of these centres offer laboratories and technical facilities.

Finally, we have verified the presence of a geographical variable around the regions of Catalonia (18 centres) and to a lesser extent Andalusia (19 centres), two outlying regions which draws together a great proportion of CSIC centres. However, this variable is not detected in the central region of Madrid, with more than forty centres. This allows us to suggest that the geographical proximity between centres affects the motivations to create links or colinks in a web environment.

References

Aguillo I. (2004). S&T, Web and Information Society Indicators: Building a new scenario. *8th Science and Technology Indicators Conference, Leiden,*

Aguillo, I. & Granadino, B. (2004). Estudio cibermétrico sobre el impacto y posición institucional del CSIC en el web. *Revista General de Información y Documentación*, 14 (2), 21-27.

Ahlgren, P., Jarneving, B. & Rousseau, R. (2003). Requirement for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology*, 54 (6), 550-560.

Bar-Ilan, J. (2002). **Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology,*** 53 (4), 308-319.

Bensman, S. J. (2004). Pearson's r and Author Cocitation Analysis: A Commentary on the Controversy. *Journal of the American Society for Information Science and Technology*, 55 (10), 935-936.

Jones, W. & Furnas, G. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science*, 36(6), 420-442.

Leydesdorff, L. (2004a). Similarity Measures, Author Cocitation Analysis, and Information Theory. *Journal of the American Society for Information Science & Technology*. Retrieved January 26, 2005 from: http://users.fmg.uva.nl/lleydesdorff/jasist04/

Leydesdorff, L. (2004b). The University-Industry Knowlege Relationship: Analyzing Patents and the Science Base of Technologies, *Journal of the American Society for Information Science and Technology,* 55(11), 991-1001. Retrieved January 26, 2005 from: http://users.fmg.uva.nl/lleydesdorff/HiddenWeb/index.htm

Leydesdorff, L. & Zaal, R. (1988). Co-words and citations relations between document sets and environments. In Egghe & Rousseau (Eds.), *Informetrics 87/88* (pp. 105–119). Amsterdam: Elsevier.

Ortega-Priego, J. L. (2003). A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of Biology and Biomedicine Centres of two European National Research Councils from a Webometric view. Scientometrics, 58 (2), 429-443. Retrieved January 26, 2005 from: http://internetlab.cindoc.csic.es/cv/11/Ortega2003.pdf

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight, *Cybermetrics*, 2,3 (1). Retrieved January 26, 2005 from:
http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. Auckland, etc.: McGraw-Hill.

Thelwall, M. (2001). The Responsiveness of Search Engine Indexes. *Cybermetrics*, 5 (1) Retrieved January 26, 2005 from:
http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking, *Journal of Documentation*, 58 (5), 563-574. Retrieved January 26, 2005 from: http://www.scit.wlv.ac.uk/%7Ecm1993/papers/2002_Existence_of_geographic_trends_jdoc.pdf

Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes, *Information Processing & Management*, 40(4), 693-707. Retrieved January 26, 2005 from: http://www.scit.wlv.ac.uk/~cm1993/papers/search_engine_bias_preprint.pdf

White, H. D. (2003). Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology*, 54 (13), 1250-1259.