doi:10.5937/BIOAC-76

Journal Classifications Based on Citation Data: The Comparison and Suitability of Three Distance Measures

Dejan Pajić

dpajic@ff.uns.ac.rs
University of Novi Sad, Faculty of Philosophy, Dr Zorana Đinđića 2, 21000 Novi Sad (Serbia)

Abstract

In the study presented in this paper, citation data from the Serbian Citation Index were used to calculate three types of proximity measures among Serbian scientific journals in the fields of social sciences and humanities. The measures were based on the frequency of intercitations among journals, journal co-citation counts, and bibliographic journal coupling. The clustering solutions derived from different distance matrices were compared and validated against the current subject classification of national journals. All three solutions were generally compatible with the list of major disciplines suggested by the Serbian Ministry of Education and Science, but the results indicated the need for a more precise subject classification. The most accurate journal classification has been achieved by using the bibliographic journal coupling method. Bibliographic journal coupling has primarily produced clusters of thematically similar journals, while intercitation and co-citation clusters also revealed some types of relationships not necessarily determined by subject similarity. The visualization of proximity data provided additional insight into the relations among disciplines and the status of several multidisciplinary journals. In general, the presented results indicate that the citation data are suitable not solely for journal classification tasks, but also as a mean of scientific domain analysis at the journal level.

Keywords: journal mapping; journal clusters; co-citation analysis; bibliographic coupling

Introduction

Monitoring and improving the quality of academic journals are among the most important issues of scientific policies, particularly those of small and developing countries. The commonly used indicators of journal quality

Journal Publishing in Developing, Transition and Emerging Countries: Proceedings of the 5th Belgrade International Open Access Conference 2012, pages 25-35. Edited by P. Šipka

(i.e *Impact Factor* and *Eigenfactor*) are usually based on the data provided by Thomson Scientific Journal Citation Reports, and are thus not suitable for the evaluation of journals from "peripheral" countries. This was the main reason why the project on the evaluation of Serbian scientific journals was launched in 2004. Since then, the Center for Evaluation in Education and Science, with the support of the Serbian Ministry of Education and Science, has annually published the *Journal Bibliometric Report (JBR)*. JBR provides quantitative indicators of impact and bibliometric quality for more than 370 Serbian academic journals. The indicators are based on the data extracted from *SCIndeks - Serbian Citation Index* (Šipka, 2005) and the *Web of Science*.

The interpretation of JBR indicators relies heavily on the appropriate definition of the reference groups of journals with similar expected (theoretical) impact. Such groups are usually formed by classifying journals into categories based on subject similarity. A generally accepted view is that the Impact Factor can be considered meaningful and valid only if journals are compared within a particular scientific field or research area (Testa & McVeigh, 2004). JBR supports two such journal disciplinary classifications: the Field of Science and Technology Classification from the Frascati Manual (FOS), and the classification provided by the authorized Field Committees of the Ministry of Education and Science of Serbia (MESS).

Generally, the problem of journal classification may be formulated as a question of whether to use *ex ante* or *ex post* classification (Leydesdorff, 2006). Journals, including those newly established, could be classified "by force" into one of the already existing groups or subject areas. Alternatively, relationships among journals could be analyzed by calculating different types of proximity measures in order to extract clusters of journals with a similar thematic scope or citation patterns. Two techniques are commonly used: co-citation analysis (Small, 1973) and bibliographic coupling (Small & Koenig, 1977). Recently, several novel methods have been introduced in order to increase the accuracy of classification and visualization of clustered journals (Janssens, Zhang, Moor & Glänzel, 2009; Boyack & Klavans, 2010).

The analyses presented in this paper were performed on a sample of journals in the fields of social sciences and humanities (SS&H). The main in-

tention was to test the suitability of journal citation data for classification tasks in those disciplines, since it is well known that SS&H are more nationally oriented and rely to a greater extent on the non-periodical literature as a principal channel of scientific communication (Hicks, 2004). Furthermore, the two JBR classifications differ significantly, particularly in the fields of SS&H. Hence, the second goal of this study was to compare the existing journal subject categories with different types of journal clusters extracted from journal citation data.

Method

Data source

All data were extracted from the SCIndeks database. The first step was to create a list of journals in the fields of SS&H. Only active journals, having published at least one issue in 2010 or 2011, were taken into account. The motivation was to classify all active journals and, hence, there were no conditions set regarding the minimal number of journal citations or co-citations among journals. The final list resulted in 137 journal titles covering different subject areas, from psychology and sociology, to history, linguistics, and literature.

The second step was to create a set of articles. In addition to the articles published in the selected journals, the sample also included all articles that cited any of the articles published in the selected journals. This means that the sample contained articles published in journals not necessarily restricted to SS&H. The publication period was limited to 10 years (2002-2011). The final sample contained 22,863 articles.

Similarity measures

Three types of similarity measures among journals were calculated. The first measure was based on *journal intercitation relations* (IC). The similarity of two journals was defined as the number of articles from the first journal that cited or had been cited by any article from the second journal. The second measure was based on *journal co-citation frequency* (CC). The linkage strength of two journals was defined as the number of articles citing both of them. Finally, the third measure was based on *bibliographic journal coupling* (BC). The strength of journal coupling is usually defined

as the number of documents cited by the two journals (Small & Koenig, 1977). However, when applied to the selected sample, this type of distance measure resulted in a sparse similarity matrix which could not be used for further analyses. As an alternative, the strength of journal coupling was defined as the number of pairs of articles from two journals citing the same journal.

Journal clustering

The resulting matrices were normalized in order to obtain more accurate similarity measures. Intercitation and co-citation matrices were normalized using the *Jaccard Index*. Since the degree of similarity in the case of bibliographic coupling was actually the number of pairs of documents, the matrix was normalized by dividing the number in each cell by the maximum number of document pairs for two journals. For example, if the number of documents for journals A and B are 50 and 60 respectively, then the actual number of pairs of documents citing the same source was divided by the value of 3000.

Normalized similarity matrices were transformed into distance matrices by calculating the inverse of the cell values ($d_{ij} = 1 - a_{ij}$). Journal clusters were derived from each distance matrix by using the *Partition Around Medoids* (PAM) algorithm in the R statistical package (Kaufman & Rousseeuw, 2008). Additionaly, *nonmetric multidimensional scaling* (MDS) was used to create journal maps and visualize the realtions among clusters.

Results

Validation of journal classifications

Social sciences and humanities journals represented in the JBR are currently classified into eight major categories according to the MESS classification: 1. economics, 2. philosophy, sociology, and political sciences, 3. psychology and educational sciences, 4. law, 5. language and literature, 6. history, archeology, and ethnology, 7. other social sciences, and 8. other humanities. The main differences between the FOS and MESS classifica-

tions are that the first one treats the fields of psychology and educational sciences as separate clusters, andalso philosophy is in the category of humanities. Both classifications were used as external criteria for the validation of different clusterings based on citation data. The *Adjusted Rand Index* (ARI) was used as a measure of similarity among five clusterings. Generally, ARI values can range from 0 to 1, where the value of 1 indicates that two classifications are exactly the same.

The ARI values are shown in Table 1. They are relatively low, but it should be noted that this does not necessarily indicate a low similarity among classifications nor a poor clustering, since a large number of journals naturally belong to more than one category. All classifications based on citation data are more similar to the MESS than to the FOS classification. The clustering based on the BC method showed the strongest agreement with both JBR classifications.

Table 1. Values of the Adjusted Rand Index of similarity among five journal clusterings

	intercitations (IC)	co-citations (CC)	bibliographic coupling (BC)
MESS	**0.43**	**0.34**	**0.53**
FOS	**0.36**	**0.30**	**0.37**
IC		0.58	0.60
CC			0.49

MESS and FOS are current journal classifications used in the Journal Bibliometric Report

Average silhouette widths (ASW) were used as measures of cluster validity and, indirectly, as an estimate of the relative suitability of different distance measures. ASW values, as well as the silhouette values of each individual object, can range from -1 to 1. Larger values indicate stronger cluster structure and greater similarity among cluster members. The classification based on the IC matrix had the lowest ASW (0.10). However, even in this case, the clustering algorithm did manage to reproduce the overall structure of JBR classification, since the six major categories of journals were clearly present. The remaining two clusters were composed

of journals in the field of political sciences, and journals in the field of physical education. All eight clusters had at least two members with negative values of silhouette, which could indicate that those journals belong to another (or independent) cluster. The clusters of economics, history, and sociology had the largest number of such spurious memberships.

The clustering solution based on journal co-citation frequencies also had relatively low ASW (0.14). The extracted clusters were very similar to those generated from the IC matrix in terms of subject contents. The political sciences cluster was also present, but sport sciences journals were merged with the field of psychology and educational sciences. In contrast, this procedure isolated a cluster of journals in the field of arts. When compared to the IC solution, it was noticeable that a large number of journals changed their cluster membership simply by switching positions between related disciplines, e.g. from sociology to political sciences, from economy to law, or from law to political sciences. However, there were more than twenty journals which were obviously completely randomly assigned to clusters. A further inspection revealed that those were the journals that had very low citation rates and, consequently, weak or no co-citation linkages with other journals.

Finally, the clustering solution based on the bibliographic journal coupling method yielded the most reasonable classification of journals, with the ASW being 0.37. Six of the eight clusters were thematically the same as those extracted from the intercitations and co-citations matrices. Compared with the classifications based on IC and CC data, the use of bibliographic coupling resulted with more “natural” clusters. Clusters were much more homogeneous and all of them had ASW values above 0.30, which was rather high given the nature of citation connections and journals as objects of classification (Janssens et al., 2009). The single exception was the cluster of journals in the fields of sociology, philosophy and political sciences. However, not even this cluster could have been labelled as spurious or inappropriate. It was obvious that those journals did belong to a single cluster, but because of their strong connections with journals from other disciplines (e.g. law and psychology), the average silhouette width of this cluster was only 0.18.

Optimal number of journal clusters

All three distance measures based on citation data have proved to be a solid basis for the separation of journals from the six major disciplines listed above. Low ASW values and the occurrence of several clusters of highly specialized journals could have been indicators that the division into eight clusters was not sufficient to provide an acceptable differentiation of disciplines. However, increasing the number of clusters in the PAM model did not improve the homogeneity of individual clusters, nor the overall quality of clustering. Moreover, in the case of bibliographic coupling, after setting the number of clusters to higher values, ASW constantly decreased. In the case of distance measures based on intercitation and co-citation data, increasing the number of clusters did result with higher values of ASW, but such improvement was obviously due to a high homogeneity of several clusters consisting of only three or four members. Additionally, at least two clusters based on intercitation data have apparently emerged as a result of the regional proximity of journal publishers.

ASW values for all three solutions became more or less constant for the number of clusters larger than 13, indirectly suggesting the optimal number of journal categories. Since it was rather difficult to assess the feasibility of clusters without broader insight into the relations among disciplines, the clustering results were visualized using nonmetric MDS. Only the maps based on bibliographic coupling and co-citation data are presented in this paper. Except for the lack of space, the reasons for such a decision were the high similarity between IC and CC maps, and the presumed higher objectivity of co-citation counts as a proximity measure. Initially, both maps were created for all 137 journals. However, the CC map suffered from several artifacts, so in that case the sample was limited to journals having at least ten co-citations with other journals. This requirement has reduced the number of displayed journals to 124. Figures 1 and 2 show the maps of journals based on the two clustering criteria. Clusters are marked with gray circles. The most representative members of each cluster (so called *medoids*) are displayed in larger font. Some journal titles are omitted in order to make the maps more readable.

The two MDS maps have very similar structure. The “core” of social sciences is positioned centrally. It contains journals in the fields of sociolo-

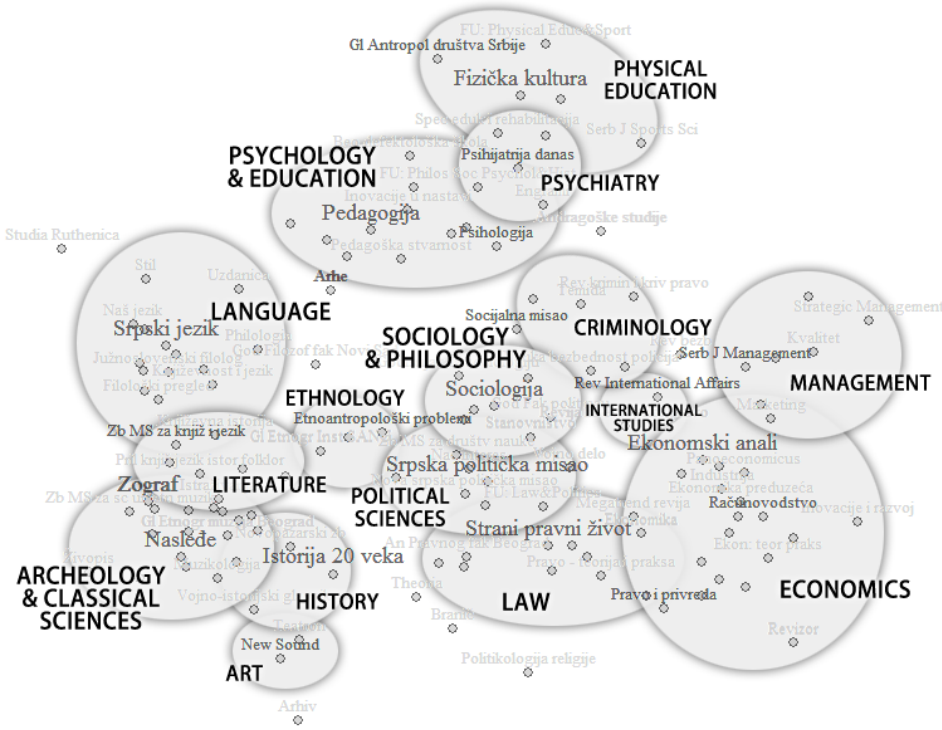

Figure 1. MDS map of 137 Serbian social sciences and humanities journals based on bibliographic journal coupling data (2002-2011)

gy, philosophy, law, history, economics, and political sciences. History and ethnology act as intermediaries or links between social sciences on one side, and humanities on the other. Both maps indicate the existence of a rather isolated cluster of disciplines obviously gravitating towards natural sciences (psychology, psychiatry, and physical education). This result is consistent with some previous findings regarding the position of psychology as a "hard" social science discipline (Ding, Chowdhury & Foo, 2000). The bibliographic coupling method has yielded clusters that were more homogeneous and compact. This was primarily due to a denser matrix when compared to the matrix of co-citations. However, it could be assumed that bibliographic coupling is probably a more appropriate and subtle technique for detecting subject similarities.

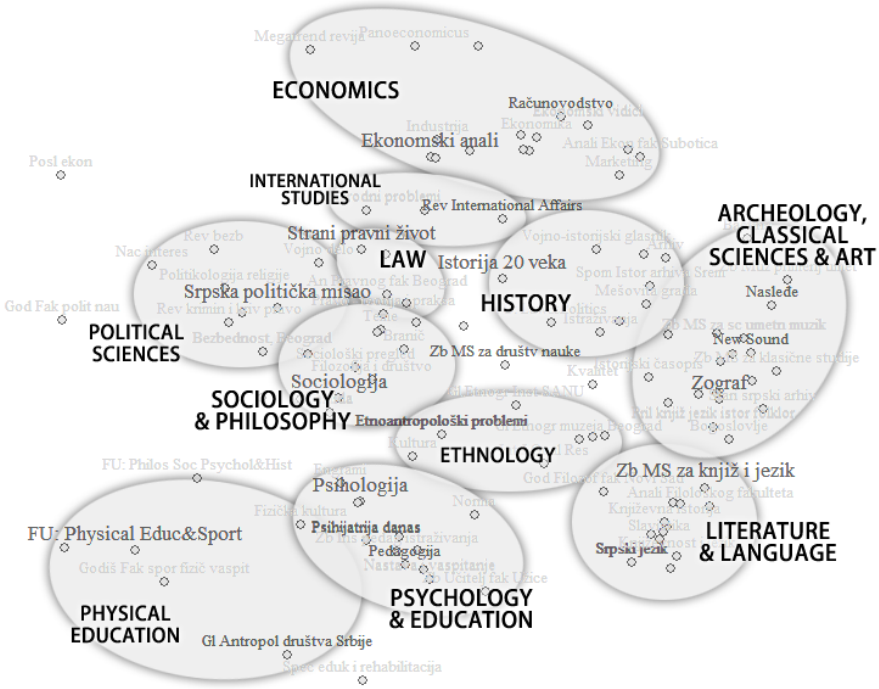

Figure 2. MDS map of 124 Serbian social sciences and humanities journals based on journal co-citation data (2002-2011)

Individual journal positions

The described methodology has revealed several "migrations" of journals among disciplines. For example, on the basis of co-citation patterns, the *Psihologija (Psychology)* journal is classified together with educational sciences journals, but on the basis of the sources used (BC) it is closer to psychiatry. Similar differences can be observed in the case of several history journals. Furthermore, it seems that some journals are far from their "parent" clusters. For example, the *Journal of the Anthropological Society of Serbia (Glasnik antropološkog društva Srbije)* is closer to sports sciences than to its current reference group of mainly ethnology journals. Finally, both maps revealed the central positions of several multidisciplinary journals.

Conclusions

The presented results have indicated that journal citation data are suitable not only for classification tasks in the fields of SS&H, but also as a mean of scientific domain analysis at the journal level. Bibliographic journal coupling has served as a good and robust alternative to co-citation and intercitation analysis, and has confirmed some previous findings regarding the accuracy of research front representation using the BC method (Boyack & Klavans, 2010).

The answer to the question about the optimal number of journal clusters cannot be simple and precise. The results have shown that the current MESS journal classification is valid, but should be more specific. However, increasing the number of groups could split up some heterogeneous but "natural" clusters, since there are many multidisciplinary journals. In that sense, it should be very useful to repeat this analysis on the sample of all national journals.

Finally, some previous research results have suggested that field normalization should solve the problem of short citation windows (van Leeuwen, 2006). Although this analysis was based on data covering a ten-year period, the resulting matrices were very sparse, and yielded several serious artefacts. Additional analyses for different and shorter periods of time would be very useful to shed more light on this problem.

References

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *JASIST*, 61, 2389-2404. doi:10.1002/asi.21419.

Ding, Y., Chowdhury,G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of Information Retrieval area, 1987-1997. *Scientometrics*, 47, 55-73. Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-540-45175-4_45

Hicks, D. (2004). The four literatures of social science. In H.F. Moed, W. Glänzel & U. Schmock (ed.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 473-496). New York: Kluwer Academic Publishers. Retrieved from http://link.springer.com/chapter/10.1007%2F1-4020-2755-9_22

Janssens, F., Zhang, L., Moor, B. D., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Inform. Processing & Management, 45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003.

Kaufman, L. & Rousseeuw, P. J. (2008) *Finding groups in data: An introduction to cluster analysis*, Hoboken: John Wiley & Sons.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *JASIST, 57*(5), 601-613. doi:10.1002/asi.20322.

Small, H. (1973). Cocitation 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. doi:10.1002/asi.4630240406.

Small, H. G., & Koenig, M. E. D. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management, 13*(5), 277-288. doi:10.1016/0306-4573(77)90017-6.

Šipka, P. (2005). The Serbian Citation Index: Context and content. In; *Proc. of ISSI 2005 - 10th Int. Conf. of the SSI*, Stockholm, Sweden, July 24-28, 2005 (pp. 710-711). Stockholm: ISSI and KUP. Retrieved from http://ceon.rs/pdf/Sipka_SCIndeks_proceedings.pdf

Testa, J. & McVeigh, M.E. (2004) The impact of open access journals: A citation study from Thomson ISI. Retrieved from http://www.lib.uiowa.edu/scholarly/documents/ISI_impact-oa-journals.pdf, 2012 March 23.

van Leeuwen, T. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics, 66*(1), 133-154. doi:10.1007/s11192-006-0010-7.