Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research

Zhang Lin[1,2], Wolfgang Glänzel[1,3], Frizo Janssens[1,4,5]

[1]K.U. Leuven, Steunpunt O&O Indicatoren, Dept. MSI, Leuven, Belgium.
[2]WISE Lab, Dalian University of Technology, Dalian, China
[3]Hungarian Academy of Sciences, IRPS, Budapest, Hungary.
[4]K.U. Leuven, ESAT-SCD, Leuven, Belgium
[5]Attentio NV, M-Brussels Village, Brussels, Belgium

1 Introduction

The history of cognitive mapping of science is as long as the history of computerised Scientometrics itself. While the first visualisations of the structure of science were considered part of information services, i.e., an extension of scientific review literature (Garfield, 1988), bibliometricians soon recognised the potential value of structural science studies for science policy and research evaluation as well. At present the identification of emerging and converging fields and the improvement of subject delineation are in the foreground.

The main bibliometric techniques are characterised by three major approaches, particularly the analysis of citation links (cross-citations, bibliographic coupling, co-citations), the lexical approach (text mining), and their combination. The sudden, large interest these techniques have found in the community is contrasted by objections and criticism (e.g., Noyons, 2001, Jarneving, 2005). Nonetheless, cognitive maps are useful tools in visualising the structure of science and can be used to adjust existing subject classification schemes even on the large scale as we will demonstrate in this study.

Clustering based on co-citation and bibliographic coupling has to cope with several severe methodological problems. This has been reported, among others by Hicks (1987) in the context or co-citation analysis and by Janssens et al. (2008) and Jarneving (2005 with regard to bibliographic coupling. One solution is to combine these techniques with other methods such as lexical-based approaches (Braam et al., 1991), (Janssens et al., 2007, 2008), or to make use of direct reference-citation links among pre-defined units as, for instance, journal cross-citations. This solution has been used by Leydesdorff (2006) and Leydesdorff and Rafols (2008).

In the present paper we use a similar approach, but in contrast to the method applied by Leydesdorff we will not base the analysis on the Journal Citation Reports which would confine us to data as available in the JCR. By contrast we calculate citations on a paper-by-paper basis and then assign individual papers to the journals in which they have been published. This offers four important opportunities. 1. We can select document types, 2. we can use pre-defined publication periods and citation windows, 3. we can use keywords extracted from the papers to characterise the cognitive composition of the clusters and 4. we can combine the citation-based classification with other methods, e.g., with text mining. In this paper we present the results of the cross-citation analysis using the first three options and we compare these results with the structure of an existing "intellectual" subject classification scheme. The aim of this comparison is exploring the possibility of using the results of the cluster analysis to improve the subject classification scheme in question.

2 Methods and results

All papers of the type article, letter, note and review indexed in the ISI Web of Science in the five-year period 2002-2006 have been taken into consideration. Citations to individual papers have been aggregated from the publication year till 2006. The complete database has been indexed and all cognitive terms extracted from title, abstract and keywords have been used for "labelling" the ob-

tained clusters. The 15-field subject classification scheme of the Steunpunt O&O Indicatoren (SOOI) developed by GLÄNZEL and SCHUBERT (2003) is used as the "control structure".

The analysis is conducted in the following five steps.

1. Studying the cognitive structure based on cross-citation cluster analysis
2. Evaluation of classification according to the cluster analysis
3. Labelling clusters using most relevant terms
4. Comparison of subject and cluster structure
5. "Migration" of journals among subject fields.

For the cluster analysis we use the agglomerative hierarchical clustering algorithm with Ward's method. In the first step, an optimum number of clusters was searched for.

In particular, we have found 15 an appropriate number of clusters. This coincides with the number of major fields according to the Leuven/Budapest scheme (12 science fields and 3 fields in the social sciences and humanities).

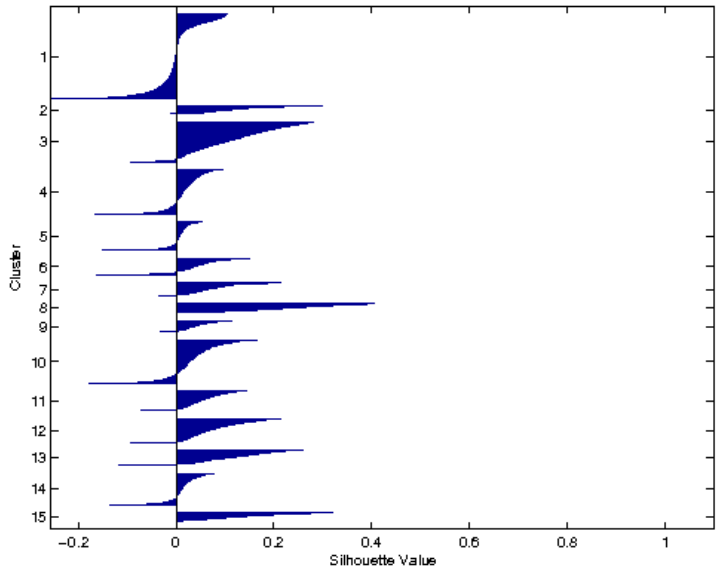


FIGURE 1. Silhouette profiles of the fifteen clusters obtained from cross-citation clustering

The silhouette values (cf. Figure 1), which express the contrast between intra- and inter-cluster similarities, substantiate that the results of this analysis are acceptable except for cluster #1. Consequently, this cluster might represent a somewhat heterogeneous, possibly less consistent subject category. The best 50 TF-IDF terms characterising the clusters (not shown) confirmed this observation. Summarising the results, we obtain the following structure (see Figure 2): biosciences (#3), neurosciences (#13), two medical clusters (#7 and #10), agriculture and environment (#9), biology (#6 and #11), geosciences (#2), chemistry (#4), physics (#12), engineering and computer science (#1), mathematics (#8), economics (#15) and two further social sciences clusters (#5 and #14). The cognitive structure of the cross-citation clusters clearly corresponds to the SOOI scheme, however, we find some interesting deviations as well. The cross-citation networks among clusters and SOOI fields, respectively, visualise similarities and differences among the two structures. For the visualisation presented in Figure 2 we used Pajek (BATAGELJ and MRVAR (2002)). Finally, the comparison of the individual journals' assignment to clusters and SOOI fields shows the "journal migration", that is, the journal's possible affinity with a *different* field. Together with the structure according to the clustering this information can be used to adjust the underlying intellectual classification scheme.

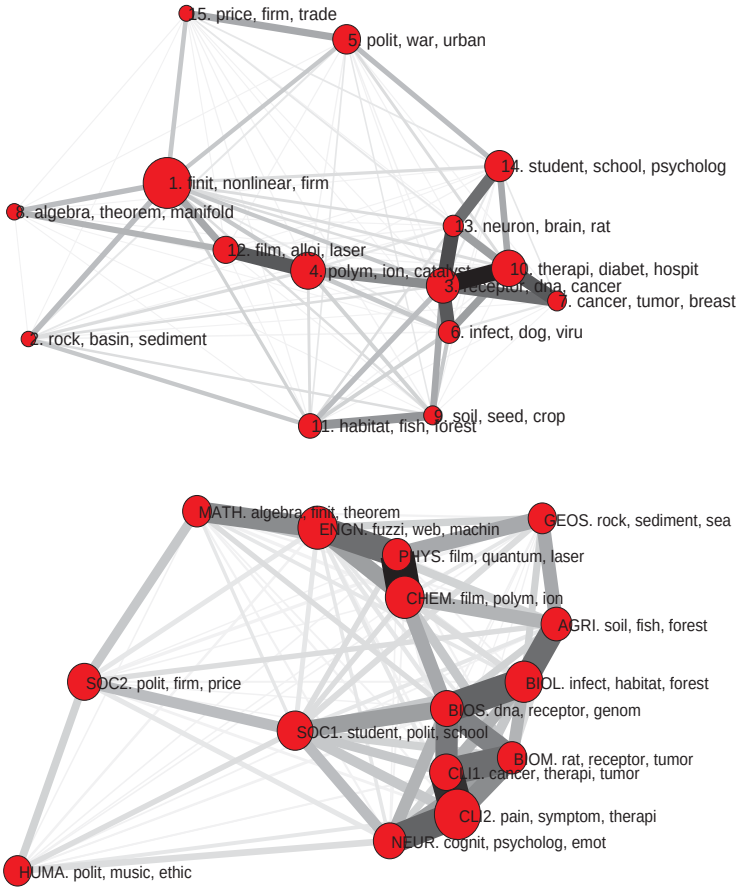


FIGURE 2: Cross-citation networks among clusters (top) and SOOI fields[1] (bottom) represented by the three most important TF-IDF terms. Since the SOOI scheme is not a partition, i.e., journals can be assigned to more than one field, more cross-citations (thicker, darker lines) can be observed among fields in the lower figure

3 Conclusions

The hard-clustering algorithm of the journal cross-citation analysis provides important information for the improvement of the SOOI scheme even if the latter one does not form a partition since it allows multiple assignment of journals to fields. We expect that the combination with a lexical approach (JANSSENS et al., 2008) can improve the efficiency of classification, but this will be part of future research.

References

Batagelj, V., Mrvar, A. (2002), Pajek - analysis and visualization of large networks, Graph Drawing, 2265: 477-478.

Braam, R.R., Moed, H.F, van Raan, A.F.J. (1991), Mapping science by combined co-citation and word analysis 1: Structural aspects, Journal of the American Society for Information Science, 42 (4): 233-251.

Garfield, E. (1988), The encyclopedic ISI Atlas of Science launches three new sections - biochemistry, immunology, and animal & plant sciences, Current Contents, 1988 (7) 3-8.

Glänzel, W., Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, Scientometrics, 56 (3): 357-367.

[1] Abbreviations of SOOI fields are as follows. AGRI = Agriculture & Environment; BIOL = Biology; BIOS = Biosciences; BIOM = Biomedical research; CLI1 = Clinical and experimental medicine I; CLI2 = Clinical and experimental medicine II; NEUR = Neuroscience & Behaviour; CHEM = Chemistry; PHYS = Physics; GEOS = Geosciences & Space sciences; ENGN = Engineering; MATH = Mathematics; SOC1 = Social sciences I; SOC2 = Social Sciences II; HUMA = Arts & Humanities.

Hicks, D. (1987), Limitations of co-citation analysis as a tool for science policy, Social Studies of Science, 17: 295-316.

Janssens, F. (2007), Clustering of scientific fields by integrating text mining and bibliometrics, Ph.D. thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium, http://hdl.handle.net/1979/847.

Janssens, F., Glänzel, W., De Moor, B. (2008), A hybrid mapping of information science, Scientometrics, 75 (3): 607-631.

Jarneving, B. (2005), The combined application of bibliographic coupling and the complete link cluster method in bibliometric science mapping, PhD Thesis, University College of Borås/Göteborg University, Sweden.

Leydesdorff, L. (2006), Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? Journal of the American Society for Information Science & Technology, 57 (5): 601-613.

Leydesdorff, L., Rafols, I. (2008), A global map of science based on the ISI subject categories, Journal of the American Society for Information Science and Technology, forthcoming.

Noyons E. (2001), Bibliometric mapping of science in a science policy context, Scientometrics, 50 (1): 83-98.