An Experimental Study on the Dynamic Evolution of Core Documents

Lin Zhang[1], Wolfgang Glänzel[2], Fred Y. Ye[3]

[1] *zhanglin_1117@126.com*
[1]Dept. Management and Economics, North China University of Water Conservancy and Electric Power, Zhengzhou (China)

[2] *Wolfgang.Glanzel@kuleuven.be*
[2]Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

[3] *yye@nju.edu.cn*
[3]School of Information Management, Nanjing University, Nanjing 210023, (China)
Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023 (China)

Introduction

The concept of the core of documents had originally been introduced in connection of co-citation analysis (Small 1973). The term *core documents* has later been re-introduced in the context of bibliographic coupling (BC; see Glänzel & Czerwon, 1996) and hybrid BC and text based similarities (Glänzel & Thijs, 2011) in order to identify strongly interlinked papers that form important nodes in the network of scholarly communication. In order to study stability and dynamics of core-document sets we apply two different methods to h-index related literature in the period 2005–2013 for illustration.

Data Sources and Processing

Data were retrieved from Thomson Reuters Web of Science Core Collection (WoS) following the strategy of Zhang et al. (2011), with extension of the period 2005–2013. We also added citing papers but removed duplicates and papers with less than 5 references to avoid biases in BC similarities. We obtained a final set of 3,270 documents. Figure 1 shows the annual increment of papers in this set.

Research Questions, Methods and Results

In this study we apply two different methods to determine core documents, (Method I) the traditional one according to Glänzel & Czerwon (1996) with a fixed number of links ($n = 15$) and Method II using the h-core of the network (Glänzel, 2012). In both cases we applied a *hybrid approach*. We used link strengths of 0.5 and 0.4 according to Salton's cosine measure. Using these parameters, we analysed the dynamics of core documents along the following questions.

- How is evolution of core documents reflected by the two methods?
- Do the two methods provide stable results?
- Do core documents adequately represent the evolution of the topic?

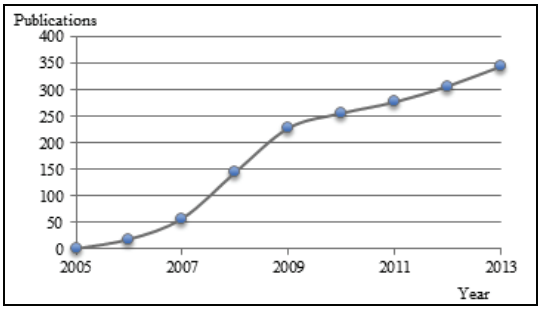

Figure 1. Distribution of h-related publications during 2005-2013.

Core document are by definition strongly interlinked with a large number of other documents in the set under study and thus represent the very core of the set. As expected, their number increases with expanding time spans, the average annual growth rate of the cumulative set amounted to 46% (Method I) and 25% (Method II), respectively. Not only the number of nodes in the network but also the number of their links is growing, however at a different pace. Indeed, we found that the complete h-related set increased at a large constant pace of 11% while the growth of the core sets was faster (see above), but its growth slowed down. This might in part be a consequence of the increasing age of references. In 2013 the core reached a representation of 2.0% and 2.4%, respectively. This characterizes the evolution of the core set with respect to the topic dynamics. The second question that arises from these figures is in how far do both methods mirror the same "core" of literature. In order to check the robustness of these methods, we compared the overlap of the sets of core documents obtained from the two methods. To this end we used BC with fixed number of links as reference standard. Concordance with Method I ranged between 83.8% and 95.2% with increasing trend from 2005–2007 to 2005–2013 and using Method II the shares ranged between 96.8% and 80.7%, however with decreasing trend.

In order to answer the third question, we analysed the core sets obtained from the two methods on the basis of authors and topics of the individual papers. The evolution of the core-document sets according to Method II is shown at three different stages in Figure 2 using Pajek with Kamada–Kawai layout (Batagelj & Mrvar 2003).

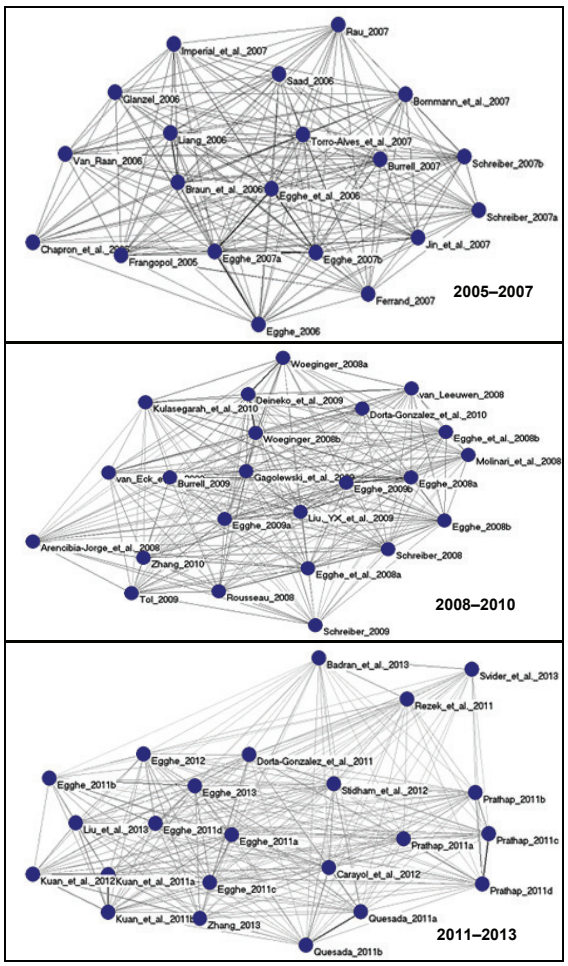

Figure 2. The evolution of the core-documents set (II).

Core nodes in Figure 2 are based on BC but hybrid similarities are used to measure the links between the nodes. This can be done because of the strong concordance between the sets obtained from the two methods. The links between core nodes in Figure 2 are denser and stronger than in the BC approach, which is due to the inclusion of textural information. The interpretation of Figure 2 is not straightforward, but the structural changes of the networks during different periods presented here are quite clear and noteworthy. The network in the first sub-period (2005–2007) comprises above all theoretical publications. The network of 2008–2010 already reflects a different picture. While most theoretical papers are still located in the centre of the network, also 'applied studies' started to appear in the core-documents set. These are distributed at the periphery of the network, which indicates that the topic starts to expand from pure theory to more application. The network of the last sub-period (2011–2013) reflects the clearest structure, where we could distinguish several sub-networks. As the most stable contributor, Egghe's six papers are found in one strongly interlinked sub-network, with the most theoretical roots. Unlike the network in 2008–2010, where some 'applied studies' were still scattered at the periphery of the network, we found more distinct sub-networks on 'applied' research in the network in 2011–2013. In this sense, core documents appear to follow the trend of the topic that is moving away from 'hard-core' informetrics towards research evaluation at different levels of aggregation and for various purposes.

Discussion and Conclusions

In the present study we focussed on 'core documents' with their evolution in publication networks using the example of a specific but nonetheless heterogeneous paper set. The two applied methods proved robust and representative. Their coverage amounted to about 2% of the topic literature, which is in line with the expectations (cf. Glänzel, 2012) but their links lead to related documents that represent a much broader coverage of the topic h-related literature.

The evolution of the core-document network represents the general tendency of shifts in topic, authors and application in an adequate manner. This gives also evidence that Hirsch-type indices have become a tool that is used also outside the informetric community.

Acknowledgments

Lin Zhang acknowledges the NFSC Grant No 71103064, Fred Ye acknowledges the NFSC Grant No 71173187 and Jiangsu Key Laboratory Fund for financial support.

References

Batagelj, V., & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77-103). Berlin: Springer.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221.

Glänzel, W. & Bart Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309.

Glänzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, 93(1), 113–123.

Small, H. (1973). Cocitation in scientific literature – new measure of relationship between 2 documents. *JASIS*, 24(4), 265-269.

Zhang, L., Bart Thijs, B. & Glänzel, W. (2011). The diffusion of H-related literature. *Journal of Informetrics*, 5(3), 583-593.