Latent Themes of Research Fields as Eigenvectors of Bibliographic Coupling Matrices

Michael Heinz,[1] Frank Havemann,[2] Oliver Mitesser,[3] and Jochen Gläser[4]

[1] *michael.heinz@rz.hu-berlin.de,* [2] *f.havemann@sciencestudies.eu*
Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft, Unter den Linden 6, D-10099 Berlin (Germany)

[3] *oliver.mitesser@gmx.de*
Technische Universität Darmstadt, Universitäts- und Landesbibliothek Darmstadt (Germany)

[4] *Jochen.Glaser@Fu-Berlin.de*
Technische Universität Berlin, Zentrum Technik und Gesellschaft (Germany)

Introduction

The possibility that new instruments for the governance of science might reduce research diversity, which is a frequent point of discussion both in policy and policy research, motivates our search for methods for measuring the diversity of research in social units such as (inter)national scientific communities or research organisations. We are currently experimenting with the extraction of latent themes from bipartite networks of papers and cited sources, for which we use a variant of latent semantic analysis (LSA) based on singular value decomposition (SVD) of paper-source matrices (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Mitesser, Heinz, Havemann, & Gläser, 2008). We argued that SVD is a preferable algorithm for this purpose, because papers and sources can be attributed to more than one theme, which is much more realistic than usual hard clustering (Mitesser et al., 2008, p. 5; cf. Janssens, Glänzel, and De Moor, 2007).

We showed that our SVD-based diversity measure is not affected by the increasing length of reference lists over time (Mitesser *et al*., 2008, p. 7) but we were hesitant to interpret the tendency towards higher entropies as trends towards higher diversity in the two fields considered. Inspecting the extracted themes, we found that the main papers of themes often have reference lists of similar length, which hinted to the possibility that the construction of latent themes is unduly affected by the length of reference lists. To avoid this influence we modified the method by normalising the paper vectors in the paper-source matrices. The time series of theme entropy calculated with paper-source matrices weighted in such a manner does no longer show an increasing trend but has peaks and troughs.

Another problem is sampling. Diversity measures of eqally sized samples can be easier compared than those of different sizes (Rousseau & Hecke, 1999; Rousseau, Van Hecke, Nijssen, & Bogaert, 1999) but scientific production in all interesting research fields is incraesing. We therefore have to draw samples of equal size from volumes with different paper numbers.

In this poster, we discuss several schemes of sampling, weighting and dimensional reduction, which we tested with our bibliography of information science research articles.

Data and Method

We use one of the data sets from our previous investigation, namely information-science papers (download from the Web of Science, document type article) from five journals in the 20-years period 1987–2006 (for details s. Mitesser *et al*., 2008).

To avoid any influence of sample sizes on diversity measures we created same-size bibliographies by drawing the first m = 500 papers (in order of appearance) from each volume. If the volume has less than m papers we enlarge the bibliography by adding papers from following years in order of appearance till we reach m papers.

Deviating from the approach in our previous paper (Mitesser *et al*., 2008), we did not decompose the unweighted affiliation matrix A but a matrix with paper vectors that are normalised by the Euclidean norm. The elements of the matrix $B = AA^{\mathrm{T}}$ are then given by the Salton cosine measure of bibliographic coupling between papers. After SVD we use the r eigenvectors u_{ik} and eigenvalues λ_k (k = 1, ..., r) of matrix B to calculate the size y_{ik}^2 of theme k in paper i according to $y_{ik}^2 = u_{ik}^2 \lambda_k$ (cf. eq. 8 in our previous paper).

In LSA the number of dimensions is further reduced below r by omitting eigenvectors which belong to small eigenvalues. This results in a lower number of extracted latent themes, which is desirable in information retrieval. It is also necessary to reduce the dimension of the theme space because in large samples a number of themes that equals the number of papers is not realistic.

Results and Discussion

For each 500-papers sample we reduced the dimension of the theme space from 500 to 100 by

omitting all themes from size ranks 101 till 500. In 2006 theme number one is *Hirsch-index* as can be seen from the titles of those nine papers where it is the biggest theme (shares ranging from 95% to 43%):

- h-index sequence and h-index matrix: Constructions and applications.
- A Hirsch-type index for journals.
- An informetric model for the Hirsch-index.
- On the h-index – A mathematical approach to a new measure of publication activity and citation impact.
- Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively.
- Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups.
- Theory and practise of the g-index.
- Is it possible to compare researchers with different scientific interests?
- An extension of the Hirsch index: Indexing scientific topics and compounds.

Many of the themes can be interpreted in this manner but we found that selecting dimensions according to the order of eigenvalues leads to inconsistencies. On the one hand we have many themes with only small contributions to all papers. On the other hand we have some relatively big themes which consists of only two papers with strong bibliographic coupling. We therefore choose another method of dimensional reduction similar to those proposed by Valle-Lisboa and Mizraji (2007). We omitted those 200 themes which have the smallest maximum shares in papers (typically all their shares in papers of the first year of each sample are less than 5%) and redistributed their paper shares among the other 300 themes (giving each paper the same weight). By this procedure we get rid of the noise of anywhere small themes and enhance all other themes apart from those of strongly coupled pairs.

We compared the time series 1986–2006 of theme entropies calculated with both methods of dimensional reduction and found qualitative agreement. Both graphs show peaks and troughs in the same years.

The number of 300 themes extracted from 500 papers is rather high. Till 1998 it exceeds even the number of papers in the first years of each sample, which we have analysed lastly. We therefore tested what happens if we further reduce the dimension of thematic space. Trials with 150 and 100 themes showed that now the leading themes became larger but also rather vague. We found many papers with maximum shares of a theme the titles of which point to themes very different from the themes of the leading papers of the theme. This negative result can be traced back to the chosen too simple method of dimensional reduction that amputates leading themes in many papers if the threshold is too high.

We nonetheless calculated time series of entropy for both theme numbers and without restricting the analysis to papers of the first year of each sample. We got diagrams which follow each change of the size of the main component displayed. We concluded that these data do not reflect diversity of research.

Our method of dimensional reduction needs further refinement in the direction outlined by Valle-Lisboa and Mizraji (2007, p. 4138–39). They sketched an algorithm that is based on the inspection of theme vectors and not of singular values.

Further investigations will test whether reducing the impact of highly cited sources by using another weighting scheme in the affiliation matrix A will give more distinct themes. We will also separate bibliometric from information retrieval themes to get a time series of one distinct research field.

References

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, & R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6), 391–407.

Janssens, F., W. Glänzel, & B. De Moor (2007). A Hybrid Mapping of Information Science. In D. Torres-Salinas and H. F. Moed (Eds.), *Proceedings of ISSI 2007*, V 1, 408–420.

Mitesser, O., Heinz, M., Havemann, F., & Gläser, J. (2008). Measuring Diversity of Research by Extracting Latent Themes from Bipartite Networks of Papers and References. In H. Kretschmer & F. Havemann (Eds.), *Proceedings of WIS 2008*, http://www.collnet.de/Berlin-2008/MitesserWIS2008mdr.pdf

Rousseau, R. & P. Van Hecke (1999). Measuring biodiversity. *Acta Biotheoretica* 47, 1–5.

Rousseau, R., P. Van Hecke, D. Nijssen, and J. Bogaert (1999). The relationship between diversity profiles, evenness and species richness based on partial ordering. *Environmental and Ecological Statistics* 6 (2), 211–223.

Valle-Lisboa, J. & E. Mizraji (2007). The uncovering of hidden structures by Latent Semantic Analysis. *Information Sciences* 177 (19), 4122–4147.