Some Comments on the Question Whether Co-occurrence Data Should Be Normalized

Ludo Waltman and Nees Jan van Eck

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2007-017-LIS
Publication	March 2007
Number of pages	6
Persistent paper URL	
Email address corresponding author	lwaltman@few.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS	
Abstract	In a recent paper in the Journal of the American Society for Information Science and Technology, Leydesdorff and Vaughan assert that raw cocitation data should be analyzed directly, without first applying a normalization like the Pearson correlation. In this report, it is argued that there is nothing wrong with the widely adopted practice of normalizing cocitation data. One of the arguments put forward by Leydesdorff and Vaughan turns out to depend crucially on incorrect multidimensional scaling maps that are due to an error in the PROXSCAL program in SPSS.
Free Keywords	Co-occurrence data, Author cocitation analysis, Normalization, Pearson correlation, Multidimensional scaling, PROXSCAL
Availability	The ERIM Report Series is distributed through the following platforms: Academic Repository at Erasmus University (DEAR), DEAR ERIM Series Portal Social Science Research Network (SSRN), SSRN ERIM Series Webpage Research Papers in Economics (REPEC), REPEC ERIM Series Webpage
Classifications	The electronic versions of the papers in the ERIM report Series contain bibliographic metadata by the following classification systems: Library of Congress Classification, (LCC) LCC Webpage Journal of Economic Literature, (JEL), JEL Webpage ACM Computing Classification System CCS Webpage Inspec Classification scheme (ICS), ICS Webpage

Some Comments on the Question Whether Co-occurrence Data Should Be Normalized

Ludo Waltman Nees Jan van Eck

Econometric Institute, Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
E-mail: {lwaltman,nvaneck}@few.eur.nl

Abstract

In a recent paper in the Journal of the American Society for Information Science and Technology, Leydesdorff and Vaughan assert that raw cocitation data should be analyzed directly, without first applying a normalization like the Pearson correlation. In this report, it is argued that there is nothing wrong with the widely adopted practice of normalizing cocitation data. One of the arguments put forward by Leydesdorff and Vaughan turns out to depend crucially on incorrect multidimensional scaling maps that are due to an error in the PROXSCAL program in SPSS.

Keywords

Co-occurrence data, author cocitation analysis, normalization, Pearson correlation, multidimensional scaling, PROXSCAL.

1 Introduction

Recently, Leydesdorff and Vaughan (hereafter LV) [3] argued that in the analysis of cocitation (or, more generally, co-occurrence) data, one should not apply a normalization, like the Pearson correlation or the cosine, to the cocitation matrix. According to LV, one should either use raw cocitation data or one should base the analysis on the asymmetrical citation matrix rather than on

the symmetrical cocitation matrix. The position taken by LV has quite far-reaching implications, since the practice of analyzing cocitation data by normalizing the cocitation matrix is widely adopted and has been used in a large number of studies. In this report, we oppose the position of LV, and we argue that there is nothing wrong with the practice of normalizing cocitation matrices. The two arguments provided by LV against this practice are both rejected by us. Although we focus our attention on author cocitation analysis, our comments apply equally well to other analyses that are based on co-occurrence data.

2 Comparison with the mapping of cities

The first argument put forward by LV says that cocitation matrices should not be normalized because such matrices contain proximity data, which is data that can be analyzed directly, without any conversion. According to LV, normalization of a cocitation matrix may distort the data in the matrix and should therefore be avoided. LV illustrate this point by providing an example in which a matrix of distances between cities is mapped using multidimensional scaling (MDS). In the example, normalization of the distance matrix does indeed distort the data. However, in our opinion there is an essential difference between mapping cities based on a distance matrix and mapping authors based on a cocitation matrix. When cities are mapped, the resulting map should reflect the distances between the cities. These distances are provided by the distance matrix. When authors are mapped, the resulting map should reflect the similarities between the authors. The cocitation matrix, however, does not directly provide these similarities. Although similarities between authors can be derived from the cocitation matrix, one generally should not simply use the number of cocitations of two authors as a measure of the author's similarity. If this approach were taken, an author who is frequently cited would on average have high similarities to other authors, whereas an author who is rarely cited would on average have low similarities to other authors. In our opinion, this does not make sense. The number of times an author is cited might be a good measure of the significance of the author's work, but it should have no effect on the extent to which the author is considered similar to other authors. In order to correct for differences in the number of times authors are cited, cocitation matrices should be normalized, for example using the Pearson correlation. The normalized cocitation data can then be used as input to MDS. We note that many cocitation studies (e.g. [4, 5]) have used the

above motivation to justify the normalization of cocitation data.

3 Mapping authors using SPSS PROXSCAL

The second argument provided by LV against the use of normalized cocitation data is of a more practical nature. LV perform an author cocitation analysis of the data studied in [1] and make a comparison between the map obtained by applying MDS to the raw cocitation matrix and the map obtained by applying MDS to the cocitation matrix normalized using the Pearson correlation. LV observe that the map based on the normalized data is less informative than the map based on the raw data, and they conclude from this that the Pearson correlation distorts cocitation data. Unfortunately, some of the MDS maps presented by LV (the Figures 5, 9, and 12 in their paper) have not been constructed correctly. This is due to an error in the PROXSCAL program in SPSS, which is the program that was used by LV to construct their maps. In SPSS version 14.0.0 (and also in some earlier versions of SPSS), the combination of similarity data and the interval transformation is handled incorrectly by PROXSCAL. This can be seen most easily by inspecting the transformation plot provided by PROXSCAL. Using SPSS version 14.0.0, we replicated the analysis performed by LV in order to check their transformation plots. We applied interval MDS to their normalized cocitation matrix (Table 9 in [1]), which resulted in a similar map as in Figure 12 in the LV paper. When we inspected the transformation plot, we observed a linear function that was either constant or increasing (depending on the choice of the initial configuration). This clearly indicates that LV present MDS maps that have not been constructed correctly, since in the case of similarity data the transformation plot should always show a decreasing function. One of the programmers of the PROXSCAL program [2] confirmed to us that the incorrect maps are caused by an error in PROXSCAL. The error can be dealt with in two ways. In SPSS version 14.0.0 (and also in some earlier versions of SPSS), rather than the interval transformation one should use a spline transformation of degree one with no interior knots. The latter transformation is equivalent to the interval transformation and works correctly. In SPSS version 14.0.1 and higher, the PROXSCAL program has been fixed and the interval transformation can be used without any problems. The corrected versions of the Figures 9 and 12 in the LV paper are displayed in the Figures 1 and 2, respectively. The map in Figure 1 is based on the Pearson correlation between authors' citation profiles

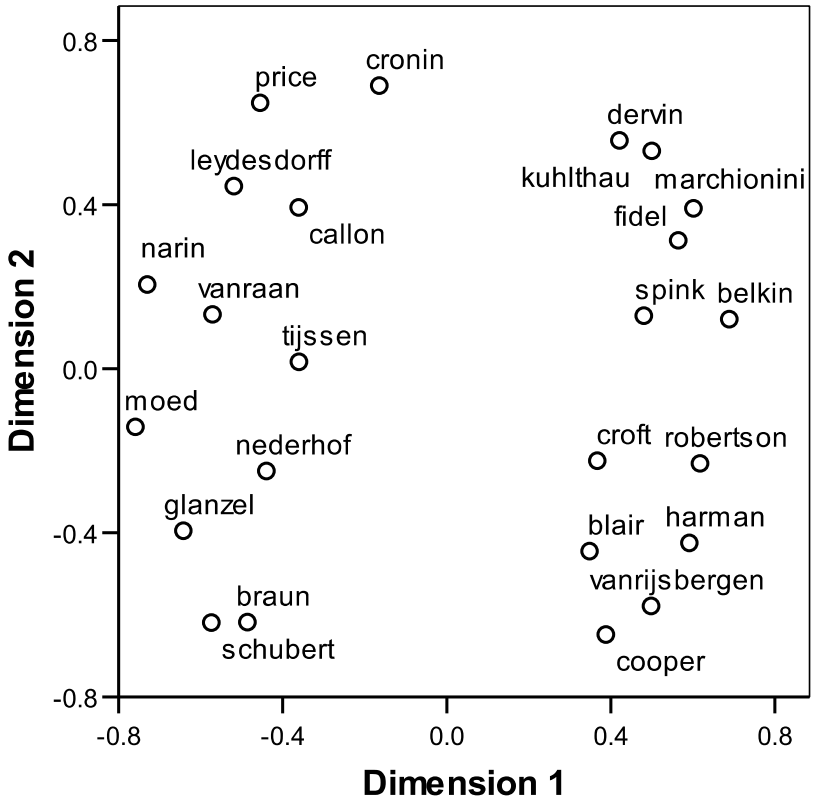

Figure 1: MDS map based on the Pearson correlation between authors' citation profiles (normalized raw stress = 0.0441). This is the corrected version of Figure 9 in the LV paper.

(Table 2 in the LV paper), while the map in Figure 2 is based on the Pearson correlation between authors' cocitation profiles (Table 9 in [1]). To reduce the effect of local minima, for each map PROXSCAL was run from ten randomly chosen initial configurations. By comparing Figure 2 in this report with Figure 11 in the LV paper, we can reconsider LV's conclusion that the Pearson correlation distorts cocitation data. The conclusion is clearly incorrect. In Figure 2, the separation between the information retrieval researchers and the scientometricians is even better than in Figure 11 in the LV paper, which seems to indicate that the use of the Pearson correlation has a positive rather than a negative effect on the quality of a cocitation map.

4 Citation data versus cocitation data

LV further argue that it is advisable to use asymmetrical citation matrices instead of symmetrical cocitation matrices as the underlying data for a cocitation map. This advice also needs to be reconsidered using the corrected MDS maps. The maps in Figure 1 in this report and Figure 8

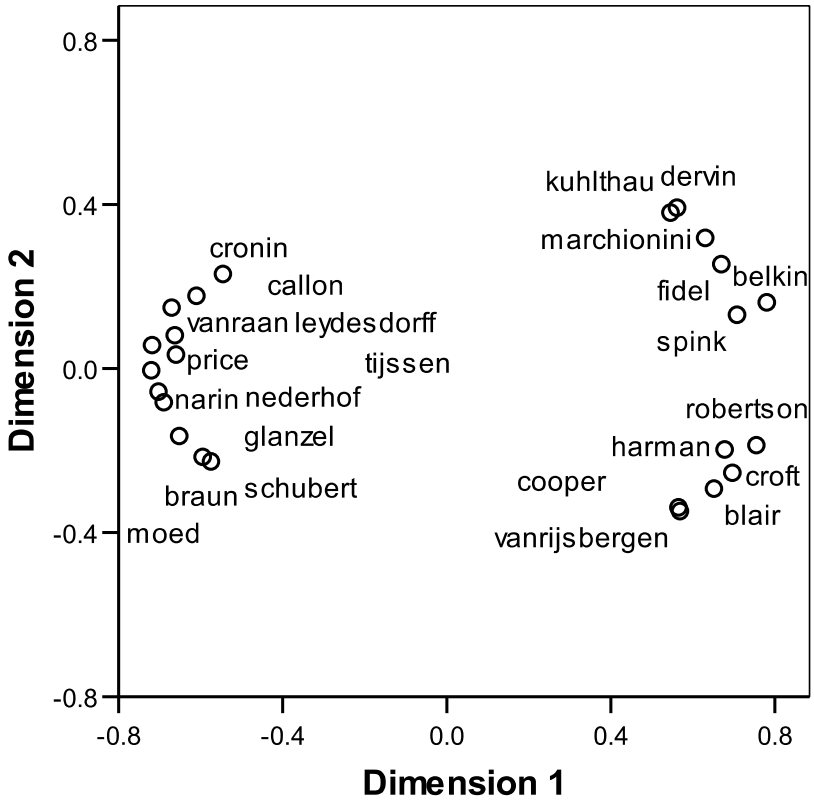

Figure 2: MDS map based on the Pearson correlation between authors' cocitation profiles (normalized raw stress = 0.0018). This is the corrected version of Figure 12 in the LV paper.

in the LV paper are based on citation data, while the maps in Figure 2 in this report and Figure 11 in the LV paper are based on cocitation data. The maps based on cocitation data show a better separation between the information retrieval researchers and the scientometricians than the maps based on citation data. So, there seems no reason to prefer citation data over cocitation data. In our opinion, more research is needed to find out whether it can be advantageous to use citation data rather than cocitation data. In addition to the Pearson correlation, other normalizations, like the cosine, could also be taken into account in such research.

References

[1] P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements 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, 2003.

[2] F. Busing. Personal communication, 2006.

[3] L. Leydesdorff and L. Vaughan. Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology*, 57(12):1616–1628, 2006.

[4] K. McCain. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6):433–443, 1990.

[5] H. White and B. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

Publications in the Report Series Research* in Management

ERIM Research Program: "Business Processes, Logistics and Information Systems"

2007

India: a Case of Fragile Wireless Service and Technology Adoption?
L-F Pau and J. Motiwalla
ERS-2007-011-LIS
http://hdl.handle.net/1765/9043

Some Comments on the Question Whether Co-occurrence Data Should Be Normalized
Ludo Waltman and Nees Jan van Eck
ERS-2007-017-LIS

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship