

# The reflection of hierarchical cluster analysis of co-occurrence matrices in SPSS

Qiuju ZHOU<sup>1</sup>, Fuhai LENG<sup>1†</sup> & Loet LEYDESDORFF<sup>2</sup>

<sup>1</sup>National Science Library, Chinese Academy of Sciences, 100190 Beijing, China

<sup>2</sup>University of Amsterdam, Amsterdam School of Communication Research (ASCoR), PO Box 15793, 1001 NG Amsterdam, the Netherlands

Received: Jul. 29, 2015

Accepted: Jul. 31, 2015

## Abstract

**Purpose:** To discuss the problems arising from hierarchical cluster analysis of co-occurrence matrices in SPSS, and the corresponding solutions.

**Design/methodology/approach:** We design different methods of using the SPSS hierarchical clustering module for co-occurrence matrices in order to compare these methods. We offer the correct syntax to deactivate the similarity algorithm for clustering analysis within the hierarchical clustering module of SPSS.

**Findings:** When one inputs co-occurrence matrices into the data editor of the SPSS hierarchical clustering module without deactivating the embedded similarity algorithm, the program calculates similarity twice, and thus distorts and overestimates the degree of similarity.

**Practical implications:** We offer the correct syntax to block the similarity algorithm for clustering analysis in the SPSS hierarchical clustering module in the case of co-occurrence matrices. This syntax enables researchers to avoid obtaining incorrect results.

**Originality/value:** This paper presents a method of editing syntax to prevent the default use of a similarity algorithm for SPSS's hierarchical clustering module. This will help researchers, especially those from China, to properly implement the co-occurrence matrix when using SPSS for hierarchical cluster analysis, in order to provide more scientific and rational results.

**Keywords** Co-occurrence matrices; Hierarchical cluster analysis; SPSS; Similarity algorithm; The syntax editor

## 1 Introduction

Co-occurrence analysis is a very important research topic within information science (e.g., co-citation, co-author, and co-word analysis). Hierarchical cluster analysis is one of the most frequently used methods to thoroughly analyze co-occurrence

<sup>†</sup> Corresponding author: Fuhai Leng (E-mail: lengfh@mail.las.ac.cn).



matrices. Statistical Product and Service Solutions (SPSS)<sup>[1]</sup> is the most widely used statistical analysis software for the hierarchical clustering of such co-occurrence matrices. But it has been found that some researchers in China rigidly “copy” the process of the hierarchical cluster analysis for co-occurrence matrices proposed by international peers, without thoroughly understanding the technical background and requirements of SPSS, and thus risk producing incorrect results<sup>[2-4]</sup>. Even more worrying is that this method has been adopted by researchers in other disciplines.

At present, only a few scholars have begun to delve into the issues arising from how one uses SPSS in hierarchical cluster analysis. Zhou<sup>[2]</sup> summarized three technical problems in clustering analysis of co-occurrence matrices: 1) preprocessing of symmetric matrices, 2) the distance among clusters, and 3) the applicability of the SPSS software. Lai<sup>[3]</sup> argued that when one chooses SPSS to do the hierarchical cluster analysis, dissimilarity matrices cannot be used as input matrices. Cui<sup>[4]</sup> stated that SPSS simply should not be used to analyze co-occurrence matrices.

Zhou<sup>[2]</sup>, Lai<sup>[3]</sup> and Cui<sup>[4]</sup> all argued that the hierarchical cluster analysis of SPSS could not be applied to co-occurrence matrices, and all three of them provided some alternative tools and methods. For example, Zhou<sup>[2]</sup> proposed multidimensional scaling analysis and social network analysis as alternatives to clustering analysis in the case of studying co-occurrence matrices. Cui<sup>[3]</sup> and Lai<sup>[4]</sup> presented in their respective blogs the solution of using occurrence matrices instead of co-occurrence matrices as input matrices; researchers in other disciplines often use the occurrence matrix. They also proposed alternative methods and tools such as R<sup>[5]</sup>, SAS<sup>[6]</sup>, MATLAB<sup>[7]</sup> or other clustering tools, and went so far as to suggest that researchers write their own programs to achieve the clustering analysis of co-occurrence matrices. Of course, the alternative tools and methods proposed by these scholars are feasible, but they do not solve the problem in SPSS directly. They simply bypassed it.

However, SPSS is the most widely used software for statistical analysis of clustering. In many cases information analysts only analyze co-occurrence matrices with SPSS, but are unwittingly using it incorrectly. The question is: Can researchers choose SPSS for clustering co-occurrence matrices? The answer is “yes,” but there has been no proper method hitherto. This paper explains the issues involved in using SPSS software for hierarchical cluster analysis of co-occurrence matrices, and proposes the corresponding solutions. The solutions are compared and tested through empirical research with the aim to guide related research.

## 2 The process of co-occurrence analysis and existing problems

The process of co-occurrence analysis includes the acquisition and preprocessing (normalization) of co-occurrence matrices, and the thorough analysis of co-occurrence



matrices, including factor analysis, cluster analysis, multidimensional scaling (MDS)<sup>[8]</sup>, social network analysis (SNA)<sup>[9]</sup>, and other visualization techniques (Fig. 1).

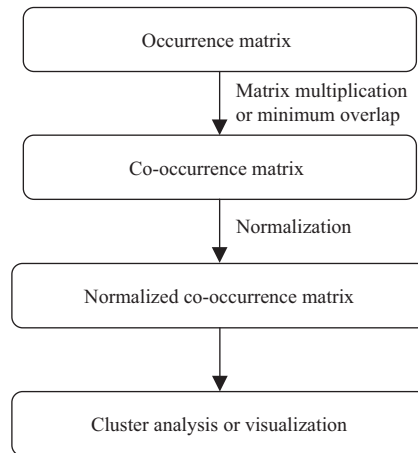


Fig. 1 The process of co-occurrence analysis.

Visualization analysis, such as MDS and SNA, use graphics to visualize the results of co-occurrence analysis. Cluster analysis evaluates relationships between the evaluated entities and classifies these entities. MDS is often used with clustering analysis<sup>[10]</sup>. Different approaches to acquisition and preprocessing of co-occurrence matrices will affect the choice of cluster analysis methods and tools. So the paper will first discuss the acquisition and pretreatment of co-occurrence matrices, and then focus on the hierarchical cluster analysis of co-occurrence matrices within SPSS.

## 2.1 Obtaining co-occurrence matrices

Occurrence matrices and co-occurrence matrices are two different forms of matrices. Occurrence matrices consist of the occurrence frequency of the category entities based on evaluated entities. They are called two-dimensional tables in SPSS or 2-mode matrices in social network analysis. Co-occurrence matrices of evaluated entities, on the other hand, usually indicate the co-occurrence frequency of evaluated entities, and can be derived from occurrence matrices. Matrix multiplication (vector inner product) and minimum overlap (intersection) are the most commonly used two methods for the derivation of the co-occurrence matrix from an occurrence



matrix<sup>[11,12]</sup>. In other words, an occurrence matrix provides the analytical basis of a co-occurrence matrix.

Much of the software in the field of information science, such as Citespace<sup>[13]</sup>, embeds the algorithms of matrix multiplication or minimum overlap to convert occurrence matrices directly into co-occurrence matrices. Researchers obtain co-occurrence matrices, but may not know the underlying derivation methods and/or the original occurrence matrix. In the Internet research (e.g. altmetrics), one usually has access only to the co-occurrence matrix of which the values are then implicitly provided using the minimum overlap method.

Using both matrix multiplication and the minimum overlap method, the co-occurrence matrices (obtained from the occurrence matrices) represent proximity matrices, in which the similarity or dissimilarity levels are indicated between the evaluated entities. For example, when one assumes keywords as evaluated entities – and, for example, documents are used as the category entities – the original keyword-document occurrence matrix can be transformed into a co-word matrix through matrix multiplication or the minimum overlap method.

## 2.2 Preprocessing of co-occurrence matrices

To date, there are three international contentions on the preprocessing of co-occurrence matrices<sup>[12,14-18]</sup>: 1) whether the Pearson correlation coefficient can be used to normalize a co-occurrence matrix, 2) how to choose the diagonal elements of a matrix, and 3) the conceptual confusion of the cosine and Ochiai index. We follow Leydesdorff & Vaughan<sup>[17]</sup>, Leydesdorff<sup>[18]</sup> and Zhou & Leydesdorff<sup>[12]</sup>, using Ochiai coefficient, Jaccard index, Dice index, equivalent index, inclusion index, etc., to normalize the co-occurrence matrix.

## 2.3 The in-depth analysis of co-occurrence matrices

Sections 2.1 and 2.2 elucidate a clear difference between occurrence matrices and co-occurrence matrices, and the methods of obtaining and normalizing co-occurrence matrices. According to the common international process for co-occurrence analysis, three approaches of multivariate analysis have been used to discern the relationships in similarity matrices: 1) factor analysis, 2) cluster analysis, and 3) MDS. These three methods are used in common statistical packages, particularly SPSS<sup>[10,15]</sup>.

The basic data in cluster analysis, MDS and social network analysis are measures of proximity between pairs of objects. But there is a trap when using SPSS for hierarchical cluster analysis. Some researchers from China fall into the trap and obtain incorrect clustering results. The following section focuses on this issue.



### 3 The problems of hierarchical cluster analysis with SPSS based on co-occurrence matrices

Hierarchical cluster analysis is the most commonly used algorithm for clustering. When obtaining the similarity matrix or distance matrix, we can evaluate the relationships between entities by means of hierarchical cluster analysis. At present, there are many tools for hierarchical cluster analysis, such as SPSS, MATLAB, R, Pajek, and UCINET. Among them, SPSS is the most commonly chosen tool because of its friendly interface and ease of use.

The process of the co-occurrence analysis<sup>[10]</sup> has been widely adopted by Chinese scholars. But many of them may be unaware of the differences between China and abroad in the habits of using the module for hierarchical clustering in SPSS. Some Chinese researchers are accustomed to using only drop down menus for the operation, but overseas scholars such as McCain<sup>[10]</sup> and White<sup>[15]</sup> usually edit syntax for the more precise operation of the software. (We are not sure whether all overseas scholars do so.)

Some Chinese researchers who “copy” the process of the co-occurrence analysis may not know that the menu options of SPSS hierarchical clustering module have embedded similarity or distance algorithms. However, the default input matrix of the module through the dropdown menu is “occurrence matrix”, and with the embedded similarity or distance algorithms, the occurrence matrix transfers to similarity or distance matrix. Some Chinese researchers input co-occurrence matrices as the input matrices and because of the recalculation of similarity inaccurate results are produced.

In fact, the menu mode of SPSS hierarchical clustering module<sup>[19]</sup> has embedded many distance algorithms, including the Euclidean distance, the squared Euclidean distance, Chebyshev distance, city block distance, Minkowski distance, and customized distance (user-defined distance), and similarity criteria such as the Pearson correlation coefficient and cosine. That is to say in the menu mode, the SPSS hierarchical cluster analysis module defaults to occurrence matrices as input matrices, and with embedded algorithm derives the proximity matrices (similarity or distance matrix), and proximity matrices are used to cluster.

International scholars such as McCain<sup>[10]</sup> and White<sup>[15]</sup> routinely edit the SPSS software’s syntax to do hierarchical cluster analysis. In the syntax editor window, without the limitations in the menu selections, researchers are able to properly choose a clustering algorithm for co-occurrence matrices. But if one inputs co-occurrence matrices into the SPSS hierarchical clustering module and operates only through menu options, without turning off the double application of a similarity or distance algorithm, one calculates the similarity twice, which results in an



overestimation of the similarity. One may wish to develop this “second-order” similarity deliberately<sup>[20]</sup>.

## 4 The empirical analysis

So far, we have explained the hazard of double application of similarity and distance algorithms embedded in the SPSS software, and we now would like to put forward appropriate solutions. The solutions come through empirical testing. Below three methods are described for clustering analysis of co-occurrence matrices, for reasons of comparison.

### 4.1 Data

We use the example of author co-citation analysis from Table 7 of Ahlgren et al.<sup>[14]</sup>. This example (see Appendix I) has been discussed previously in several contributions to the *Journal of the American Society for Information Science and Technology (JASIST)*<sup>[12,14-18,21]</sup>. This occurrence matrix is taken from Leydesdorff & Vaughan<sup>[17]</sup>. They repeated the analysis of Ahlgren et al.<sup>[14]</sup> to obtain the original (asymmetrical) data matrix. By using precisely the same searches they found 469 articles in *Scientometrics* and 494 in *JASIST* on November 18, 2004. Appendix I shows the table from Leydesdorff<sup>[18]</sup>, author co-citation matrix of 24 information scientists (Table 7 of Ahlgren et al.<sup>[14]</sup> at p. 555; main diagonal values added).

### 4.2 Methods

#### 4.2.1 Method 1

We input a (normalized) co-occurrence matrix into the hierarchical clustering module of SPSS, but only menu selections are used to operate the hierarchical cluster analysis.

- Step 1: Matrix normalization. The original occurrence matrix from Leydesdorff & Vaughan<sup>[17]</sup> was normalized using the cosine, resulting in a cosine matrix. This step can be done in the “proximities” module of SPSS. The cosine matrix is equivalent to the Ochiai index of co-occurrence matrix<sup>⊙</sup> (see Zhou & Leydesdorff<sup>[12]</sup>). However, the Pearson coefficients or cosine are not suitable to normalize co-occurrence matrices (see Leydesdorff & Vaughan<sup>[17]</sup>, Leydesdorff<sup>[18]</sup>, Zhou & Leydesdorff<sup>[12]</sup>).
- Step 2: The visualization of the cosine matrix with MDS. We input the cosine matrix into the multidimensional scaling (PROXSCAL) module of SPSS. One uses the matrix in this case as ordinal data (non-metric MDS).

<sup>⊙</sup> Within SPSS, Ochiai is only defined for the binary scale.



- Step 3: Hierarchical cluster analysis. We input the cosine matrix into the hierarchical clustering module through the menu options of SPSS.

In the data editor window of SPSS, we selected “analyze>classify>hierarchical cluster”. In the dialog box of “method”, we chose cosine in the “measure” option box. In the dialog box of “statistics”, we chose the option “proximity matrix”, and the module output is a proximity matrix which applies cosine to the input cosine matrix.

For the cluster method, we chose the “between group” method. The purpose of this paper is to compare the difference of the similarity measure being applied between an occurrence matrix and a co-occurrence matrix. So we made the similarity measures and cluster method the same across the three methods.

#### 4.2.2 Method 2

- Step 1: Repetition of the first step in Method 1.
- Step 2: Repetition of the second step in Method 1.
- Step 3: The syntax editing. We edited the syntax to prevent another round of normalization in the hierarchical clustering module of SPSS. The following steps have been taken for hierarchical cluster analysis.

First, in data editor window of SPSS, we selected “file>open>data”. We chose the cosine matrix which we obtained from Step 1, and clicked on “paste”. The syntax editor window would be opened automatically.

Second, in the syntax editor window, we selected “run>all”.

Third, in data editor window, we chose cluster analysis based on variables. In the dialog box of “method”, we chose the cosine in the “measure” option box. For the cluster method, we chose “between group” method and then clicked on “paste” (Fig. 2).

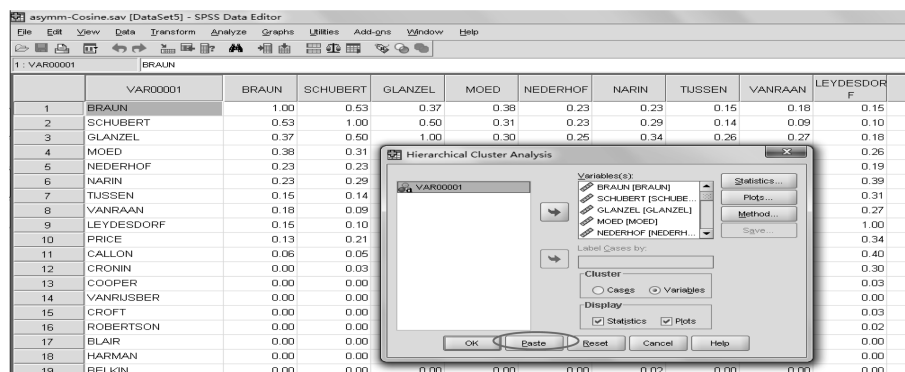


Fig. 2 Data editor window of hierarchical clustering module of SPSS.



Fourth, the syntax editor window contained the syntax of the operation in the data editor window as illustrated in Fig. 3.

```

PROXIMITIES  BRAUN SCHUBERT GLANZEL MOED NEDERHOF NARIN TIJSSEN
VANRAAN LEYDESDORF PRICE CALLON
  CRONIN COOPER VANRIJSBER CROFT ROBERTSON BLAIR HARMAN BELKIN SPINK
  FIDEL MACHIONINI KUHALTHAU DERVIN
/MATRIX OUT('C:\Users\...\Temp\spss1828\spssclus.tmp')
/VIEW=VARIABLE
/MEASURE=COSINE
/PRINT NONE
/STANDARDIZE=VARIABLE NONE.
CLUSTER
/MATRIX IN('C:\Users\...\Temp\spss1828\spssclus.tmp')
/METHOD BAVERAGE
/PRINT SCHEDULE
/PRINT DISTANCE
/PLOT DENDROGRAM VICICLE.
ERASE FILE='C:\Users\...\Temp\spss1828\spssclus.tmp'.

```

Fig. 3 The syntax of the hierarchical clustering module in SPSS.

Figure 3 shows the default syntax of the hierarchical clustering module in SPSS, which is also the same syntax as Step 3 of Method 1. Up until now, we have repeated Step 3 of Method 1, and displayed its syntax. We input the cosine matrix into the hierarchical clustering module, and applied the embedded similarity algorithm to the cosine matrix, and the syntax indicated clearly that “/MEASURE=COSINE”. If we do not prevent another round of normalization in the hierarchical clustering module in SPSS, we calculate cosine twice.

Fifth, we edited the syntax to prevent another round of normalization in the hierarchical clustering module of SPSS. In syntax editor window, we deleted the syntax “/MEASURE=COSINE”, and changed the matrix in “MATRIX IN” to the distance matrix derived from the cosine matrix which we input into SPSS.

Distance is the core concept of cluster analysis, indicating the degree of divergence between subclasses. The concept of “distance” in cluster analysis is the opposite of similarity. Therefore, we must transfer cosine matrix (similarity matrix) into a distance matrix. In this case, we changed the cosine matrix to the (1-cosine) matrix. The syntax is described as in Fig. 4.

```

CLUSTER
/MATRIX IN('C:\Users\...\matrix_in_Ahlgren 2003 1-cosine.sav')
/METHOD BAVERAGE

```

Fig. 4 The modified syntax of the hierarchical clustering module in SPSS.





### 4.2.3 Method 3

We input the occurrence matrix without pretreatment into the hierarchical clustering module of SPSS. In the menu options, the embedded similarity algorithm (here we chose cosine) was used to turn the occurrence matrix into a normalized co-occurrence matrix (cosine matrix). The resulting cosine matrix was used for MDS.

- Step 1: Cluster analysis. We input the occurrence matrix from Leydesdorff & Vaughan<sup>[17]</sup> into the hierarchical clustering module of SPSS. In the dialog box of “measure”, we chose cosine. For the cluster method, we chose “between group” method.
- Step 2: Multidimensional scaling. We input the “proximity matrix” of cosine into the multidimensional scaling (PROXSCAL) module of SPSS. One uses the matrix in this case as ordinal data.

## 5 Results and discussion

Figure 5 illustrates the dendrograms of clustering analysis from Method 1, Method 2, and Method 3. The left side of Fig. 5 shows the clustering map of Method 1 (Step 3), and the right side of Fig. 5 shows the map of Method 2 (Step 3) and Method 3 (Step 1). In other words, one obtains the same result through Method 2 and Method 3.

Method 1 is the traditional approach, as much of the literature has done. In Step 3 of Method 1, we input the cosine matrix into the clustering module, but the

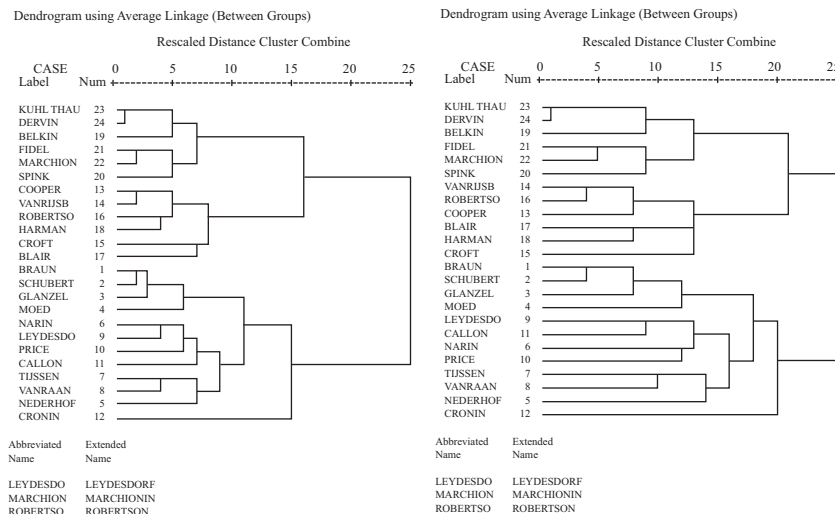


Fig. 5 Dendrograms using cluster analysis of similarity matrix of author citation (in SPSS). Note: The left graph is based on cosine of cosine (Method 1 (Step 3)) and the right graph is based on cosine (Method 2 (Step 3); Method 3 (Step 1)).



## Research Paper

embedded similarity algorithm calculated the cosine similarity one extra time. So with Method 1, we got the matrix of cosine of cosine. That is to say the cluster algorithm is based on the cosine of the cosine matrix, not on the inputting cosine matrix. So we can find that the left side is more aggregative than the right side of Fig. 5.

In Method 2, we input the same cosine matrix as Method 1 into the hierarchical clustering module in SPSS, but we prevented another round of similarity algorithm by editing the syntax. So the cluster analysis is based on the cosine matrix. By using this method we have obtained the correct clustering result.

In Method 3, we input the original occurrence matrix without any pretreatment, but changed the cluster analysis in the menu options of SPSS during the first step. We chose the cosine measure from the embedded similarity algorithm in the clustering module. So the cluster analysis is based on the cosine matrix.

The three methods can produce the same MDS map shown in Fig. 6, because they are in that case all based on the same cosine matrix as input.

In summary, Method 1 overestimated and distorted the similarity of the authors by calculating the cosine twice. But Methods 2 and 3 produced the same, correct results through different ways. For the researchers of information science, Method 2 is more applicable. This is because the input matrix of Method 2 can be a normalized co-occurrence matrix, but the input matrix of Method 3 must be an occurrence matrix.

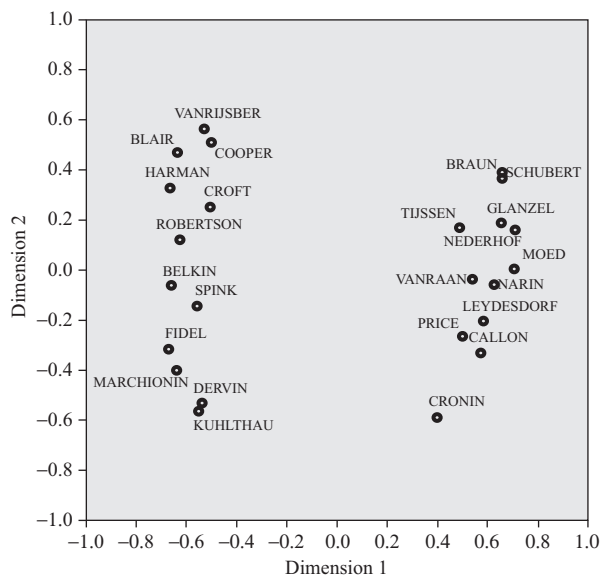


Fig. 6 MDS based on cosine (Method 1, Method 2 and Method 3).



In addition to SPSS, there are other tools such as R, SAS, MATLAB, UCINET, Pajek, which can be used to do hierarchical cluster analysis. In fact, the object of the hierarchical cluster analysis is the distance matrices. We can edit the syntax of R, SAS, and MATLAB to apply clustering algorithms to co-occurrence matrices. UCINET and Pajek can also be alternative tools to SPSS to do hierarchical cluster analysis for co-occurrence matrices.

## 6 Conclusion

This paper points out that in the menu options of the SPSS hierarchical clustering module the default input matrix is an occurrence matrix, which is converted into a normalized co-occurrence matrix through the embedded similarity or distance algorithm. But if the input is a co-occurrence matrix, one obtains a matrix in which similarity is calculated twice. The cluster algorithm based on similarities in a similarity matrix will lead to inaccurate results.

To solve this problem, this paper presents a method of editing the syntax to prevent the default use of a similarity algorithm for SPSS's hierarchical clustering module. We are hoping Chinese researchers are careful to properly implement the co-occurrence matrix when using SPSS for hierarchical cluster analysis, in order to provide more scientific and rational results.

## Author contributions

Q.J. Zhou (zhouqj@mail.las.ac.cn) performed data analyses and wrote the manuscript. F.H. Leng (lengfh@mail.las.ac.cn, corresponding author) was in charge of the overall research design, organized the discussion, proposed the research outline and revised the paper. L. Leydesdorff (loet@leydesdorff.net) proposed the research idea, revised the discussion section and edited the paper.

## References

- 1 SPSS software. Retrieved on July 29, 2015, from <http://www-01.ibm.com/software/analytics/spss/>.
- 2 Zhou, L., Yang, W., & Zhang, Y. F. Issues and re-consideration on cluster analysis in co-occurrence matrix. *Journal of Intelligence (in Chinese)*, 2014, 33(6): 32–36. Retrieved on July 29, 2015, from [http://d.wanfangdata.com.cn/Periodical\\_qbzz201406008.aspx](http://d.wanfangdata.com.cn/Periodical_qbzz201406008.aspx). DOI: 10.3969/j.issn.1002-1965.2014.06.007.
- 3 Lai, Y.G. Dissimilarity matrix and SPSS hierarchical clustering (in Chinese). Retrieved on July 29, 2015, from <http://blog.sciencenet.cn/home.php?mod=space&uid=422720&do=blog&id=313758>.
- 4 Cui, L. We cannot use SPSS to analyze co-occurrence matrix (in Chinese). Retrieved on July 29, 2015, from <http://blog.sciencenet.cn/blog-82196-328819.html>.



## Research Paper

- 5 R language definition. Retrieved on July 29, 2015, from <http://cran.r-project.org/doc/manuals/R-lang.html>.
- 6 SAS Institute Inc. SAS® 9.3 web applications: Clustering. Cary, NC: SAS Institute Inc., 2011.
- 7 MATLAB. Retrieved on July 29, 2015, from <https://en.wikipedia.org/wiki/MATLAB>.
- 8 Davison, M.L. Multidimensional scaling. New York: John Wiley and Sons, 1983.
- 9 Wasserman, S., & Faust, K. Social network analysis: Methods and applications. Cambridge: Cambridge University Press, 1994
- 10 McCain, K. W. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 1990, 41(6): 433–443. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-4571%28199009%2941:6%3C433::AID-ASII1%3E3.0.CO;2-Q/abstract>. DOI: 10.1002/(SICI)1097-4571(199009)41:6<433::AID-ASII1>3.0.CO;2-Q.
- 11 Morris, S. A. Unified mathematical treatment of complex cascaded bipartite networks: The case of collections of journal papers. Doctor dissertation. Stillwater: Oklahoma State University, 2005. Retrieved on July 29, 2015, from <http://eprints.rclis.org/6714/>.
- 12 Zhou, Q. J., & Leydesdorff, L. The normalization of occurrence and co-occurrence matrices in bibliometrics using cosine similarities and Ochiai coefficients. *Journal of the American Society for Information Science and Technology*. (To appear).
- 13 Chen, C.M. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*. 2006, 57(3): 359–377. Retrieved on July 29, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.20317/abstract;jsessionid=64AC6F2A2AD052ACD9DD0B6DA28ACBB9.f03t02>. DOI: 10.1002/asi.20317.
- 14 Ahlgren, P., Jarneving, B., & Rousseau, R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 2003, 54(6): 550–560. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.10242/abstract>. DOI: 10.1002/asi.10242.
- 15 White, H.D. Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 2003, 54(5), 423–434. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.10228/abstract>. DOI: 10.1002/asi.10228.
- 16 Bensman, S. J. Pearson's *r* and author cocitation analysis: A commentary on the controversy. *Journal of the American Society for Information Science and Technology*, 2004, 5(10): 935–936. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.20028/full>. DOI: 10.1002/asi.20028.
- 17 Leydesdorff, L., & Vaughan, L. 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*, 2006, 57(12): 1616–1628. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.20335/citedby>. DOI: 10.1002/asi.20335.
- 18 Leydesdorff, L. On the normalization and visualization of author co-citation data: Salton's cosine *versus* the Jaccard index. *Journal of the American Society for Information Science and Technology*, 2008, 59(1): 77–85. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.20732/abstract>. DOI: 10.1002/asi.20732.



- 19 Coakes, S.J., & Steed, L. SPSS: Analysis without anguish using SPSS version 14.0 for Windows. New York: John Wiley & Sons, 2009: 5–7. Retrieved on July 27, 2015, from <http://dl.acm.org/citation.cfm?id=1804538>.
- 20 Colliander, C., & Ahlgren, P. Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics*, 2012, 90(2): 675–685. Retrieved on July 27, 2015, from <http://link.springer.com/article/10.1007%2Fs11192-011-0491-x>. DOI: 10.1007/s11192-011-0491-x.
- 21 Leydesdorff, L. Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, 2005, 56(7): 769–772. Retrieved on July 27, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/asi.20130/references>. DOI: 10.1002/asi.20130.



National Science Library,  
Chinese Academy of  
Sciences

<http://www.chinalibraries.net>



National Science Library,  
Chinese Academy of  
Sciences

Appendix I: Author co-citation matrix of 24 information scientists

| Authors      | 1          | 2          | 3          | 4          | 5         | 6          | 7         | 8          | 9          | 10         | 11        | 12        | 13        | 14        | 15        | 16         | 17        | 18        | 19         | 20        | 21        | 22        | 23        | 24        | Total        |     |
|--------------|------------|------------|------------|------------|-----------|------------|-----------|------------|------------|------------|-----------|-----------|-----------|-----------|-----------|------------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|--------------|-----|
| 1            | 50         | 29         | 19         | 20         | 9         | 13         | 5         | 9          | 7          | 7          | 2         | 0         | 0         | 0         | 0         | 0          | 0         | 0         | 0          | 0         | 0         | 0         | 0         | 0         | 120          |     |
| 2            | 29         | 60         | 28         | 18         | 10        | 18         | 5         | 5          | 5          | 12         | 2         | 1         | 0         | 0         | 0         | 0          | 0         | 0         | 0          | 0         | 0         | 0         | 0         | 0         | 0            | 133 |
| 3            | 19         | 28         | 53         | 16         | 10        | 20         | 9         | 14         | 9          | 11         | 5         | 3         | 0         | 0         | 0         | 0          | 0         | 0         | 0          | 0         | 0         | 0         | 0         | 0         | 0            | 144 |
| 4            | 20         | 18         | 16         | 55         | 12        | 20         | 5         | 18         | 13         | 12         | 7         | 4         | 0         | 0         | 0         | 0          | 0         | 0         | 0          | 0         | 0         | 0         | 0         | 0         | 0            | 145 |
| 5            | 9          | 10         | 10         | 12         | 31        | 12         | 8         | 11         | 7          | 4          | 4         | 2         | 0         | 0         | 0         | 0          | 0         | 0         | 0          | 0         | 0         | 0         | 0         | 0         | 0            | 89  |
| 6            | 13         | 18         | 20         | 20         | 12        | 64         | 11        | 20         | 21         | 20         | 11        | 9         | 1         | 0         | 1         | 1          | 0         | 0         | 1          | 1         | 0         | 0         | 0         | 0         | 0            | 180 |
| 7            | 5          | 5          | 9          | 5          | 8         | 11         | 22        | 13         | 10         | 5          | 6         | 1         | 0         | 1         | 2         | 1          | 0         | 0         | 0          | 1         | 0         | 0         | 0         | 0         | 0            | 83  |
| 8            | 9          | 5          | 14         | 18         | 11        | 20         | 13        | 50         | 13         | 12         | 12        | 6         | 2         | 1         | 2         | 1          | 0         | 0         | 0          | 1         | 0         | 0         | 0         | 0         | 0            | 140 |
| 9            | 7          | 5          | 9          | 13         | 7         | 21         | 10        | 13         | 46         | 17         | 14        | 10        | 1         | 0         | 1         | 1          | 0         | 0         | 0          | 2         | 0         | 0         | 0         | 0         | 0            | 131 |
| 10           | 7          | 12         | 11         | 12         | 4         | 20         | 5         | 12         | 17         | 54         | 10        | 9         | 1         | 1         | 1         | 1          | 0         | 0         | 2          | 0         | 1         | 0         | 1         | 0         | 2            | 129 |
| 11           | 2          | 2          | 5          | 7          | 4         | 11         | 6         | 12         | 14         | 10         | 26        | 4         | 0         | 0         | 1         | 1          | 0         | 0         | 0          | 1         | 0         | 0         | 0         | 0         | 0            | 80  |
| 12           | 0          | 1          | 3          | 4          | 2         | 9          | 1         | 6          | 10         | 9          | 4         | 24        | 1         | 0         | 0         | 1          | 0         | 0         | 0          | 1         | 0         | 1         | 0         | 1         | 1            | 55  |
| 13           | 0          | 0          | 0          | 0          | 0         | 1          | 0         | 2          | 1          | 1          | 0         | 1         | 30        | 15        | 5         | 12         | 5         | 10        | 7          | 2         | 0         | 2         | 1         | 1         | 1            | 66  |
| 14           | 0          | 0          | 0          | 0          | 0         | 0          | 1         | 1          | 0          | 1          | 0         | 0         | 15        | 30        | 7         | 17         | 5         | 13        | 5          | 3         | 1         | 0         | 1         | 1         | 1            | 71  |
| 15           | 0          | 0          | 0          | 0          | 0         | 1          | 2         | 2          | 1          | 1          | 1         | 0         | 5         | 7         | 18        | 9          | 6         | 7         | 8          | 6         | 2         | 1         | 2         | 2         | 2            | 63  |
| 16           | 0          | 0          | 0          | 0          | 0         | 1          | 1         | 1          | 1          | 1          | 1         | 12        | 17        | 9         | 36        | 7          | 13        | 12        | 10         | 8         | 6         | 4         | 4         | 4         | 4            | 109 |
| 17           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 0          | 0         | 0         | 5         | 5         | 6         | 7          | 18        | 10        | 4          | 2         | 2         | 2         | 0         | 0         | 0            | 43  |
| 18           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 0          | 0         | 0         | 10        | 13        | 7         | 13         | 10        | 31        | 9          | 5         | 5         | 3         | 1         | 1         | 1            | 77  |
| 19           | 0          | 0          | 0          | 0          | 0         | 1          | 0         | 0          | 0          | 2          | 0         | 0         | 7         | 5         | 8         | 12         | 4         | 9         | 36         | 9         | 9         | 10        | 14        | 10        | 10           | 100 |
| 20           | 0          | 0          | 0          | 0          | 0         | 1          | 1         | 1          | 2          | 0          | 1         | 1         | 2         | 3         | 6         | 10         | 2         | 5         | 9          | 21        | 11        | 7         | 5         | 4         | 4            | 71  |
| 21           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 1          | 0         | 0         | 0         | 1         | 2         | 8          | 2         | 5         | 9          | 11        | 23        | 12        | 10        | 6         | 6            | 67  |
| 22           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 0          | 0         | 1         | 2         | 0         | 1         | 6          | 2         | 3         | 10         | 7         | 12        | 24        | 11        | 5         | 60           |     |
| 23           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 1          | 0         | 1         | 1         | 1         | 2         | 4          | 0         | 1         | 14         | 5         | 10        | 11        | 26        | 14        | 65           |     |
| 24           | 0          | 0          | 0          | 0          | 0         | 0          | 0         | 0          | 0          | 2          | 0         | 1         | 1         | 1         | 2         | 4          | 0         | 1         | 10         | 4         | 6         | 5         | 14        | 20        | 51           |     |
| <b>Total</b> | <b>120</b> | <b>133</b> | <b>144</b> | <b>145</b> | <b>89</b> | <b>180</b> | <b>83</b> | <b>140</b> | <b>131</b> | <b>129</b> | <b>80</b> | <b>55</b> | <b>66</b> | <b>71</b> | <b>63</b> | <b>109</b> | <b>43</b> | <b>77</b> | <b>100</b> | <b>71</b> | <b>67</b> | <b>60</b> | <b>65</b> | <b>51</b> | <b>2,272</b> |     |

Note: 1: Braun; 2: Schubert; 3: Glanzel; 4: Moed; 5: Nederhof; 6: Narin; 7: Tijssen; 8: VanRaaij; 9: Leydesdorff; 10: Price; 11: Callon; 12: Cronin; 13: Cooper; 14: Vanrijbergen; 15: Croft; 16: Robertson; 17: Blair; 18: Harman; 19: Belkin; 20: Spink; 21: Fidei; 22: Marchionini; 23: Kuhthau; 24: Dervin. We have referred to Table 7 of Ahlgren et al.<sup>[14]</sup> at p. 555; main diagonal values were added by Leydesdorff & Vaughan<sup>[17]</sup>; see Leydesdorff<sup>[18]</sup> at p. 78.