Methods # 212

Author cocitation analysis of articles from "NeuroImage"

Finn Årup Nielsen and Lars Kai Hansen
Informatics and Mathematical Modelling, Technical University of Denmark
International Neuroimaging Consortium

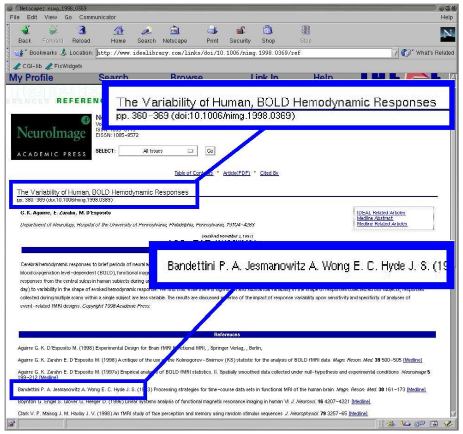
Introduction

Author cocitation analysis describes a scientific field by analyzing and visualizing the citation pattern in scientific articles [4, 5, 6]. Often ACA is based on data collected by the Institute of Scientific Information among a limited set of key authors within a field. Here we work on data from a *single journal* (the journal "NeuroImage") downloaded from the publisher website and working on an unrestricted number of authors (i.e., as many as cited). Our method is simple metric multidimensional scaling, implemented via singular value decomposition. By analyzing the titles of the articles we are able to automatically label the axes by combining the citation analysis with latent semantic indexing [1]. We found the major axis in the citation pattern of *NeuroImage* to be associated with fMRI-methodology. Other axes were associated with "memory" in one end and "motor"/"activation" in the other. A further axis showed a "statistical"/"analysis" dimension. The method allows the novice researcher to get an overview of the subfields present in a journal as well overall key authors and subfield key authors.

Data

325 "References" webpages were downloaded from the website of Academic Press (Idealibrary) covering the period (1997 February – 2000 June). These webpages contain bibliographic information for one scientific article (a document) including the citations (to other scientific documents).

The title of each of the 325 documents were extracted together with the all authors in each reference in the documents.



The individual words in the title were found excluding *stop words* (common words such as: "the", "and", ...) and converted to lower case, e.g., the title "The Variability of Human BOLD Hemodynamic Responses" is converted to {variability, human, bold, hemodynamic, responses}. The set of words from all documents were then used in the construction of an indicator datamatrix $\mathbf{Y}(N \times Q)$ containing N rows of documents and Q columns of words, where $y_{nq} = 1$ if the q'th word appears in the title of the n'th document, zero otherwise.
Similarly were all authors in the references extracted and stemmed to surname and first initial, thus "Bandettini P. A." would be converted to "Bandettini P". An indicator datamatrix $\mathbf{X}(N \times P)$ were constructed where the P columns each represents a cited author (name).

Mathematical modeling

We performed singular value decomposition (SVD) into K components on the document × cited author matrix $\mathbf{X}(N \times P)$

$$\mathbf{U}\mathbf{L}\mathbf{V}^{\mathsf{T}} = \mathrm{svd}_K\left[\mathbf{X}\right]. \tag{1}$$

We call the K vectors in:

- $\mathbf{U}(N \times K)$ for the *eigendocuments*

and similarly the K vectors in:

- $\mathbf{V}(P \times K)$ for the *eigenauthors*

As we did not extract the mean of the $\mathbf{X}$ matrix the first principal component (eigenauthor) will correspond to "authoritative sources" as defined by [2]:

- *Authorities* (authoritative authors) are the authors that score high when projected onto the first eigenauthor $\mathbf{v}_1$.
- *Hubs* (hub documents) are documents that score high when projected onto the first eigendocument $\mathbf{u}_1$.

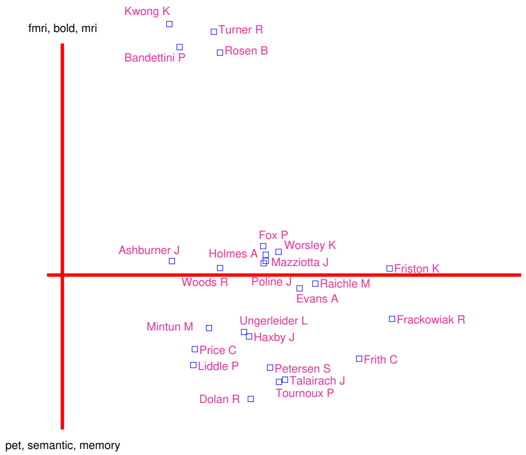
Authorities are authors that are cited by many *hub* documents and hub documents are documents that cite many authoritative authors. The authors can be visualized in two dimensions by selecting 2 eigenauthors and plotting the elements.
The axis labeling are found by a projection using the eigendocuments:

$$\mathbf{Z} = \mathbf{Y}^{\mathsf{T}}\mathbf{U}. \tag{2}$$

For the labeling of the k'th axis the 3 highest scoring words are selected from the k'th column vector $\mathbf{z}_k$ of the matrix $\mathbf{Z}(Q \times K)$.

Results

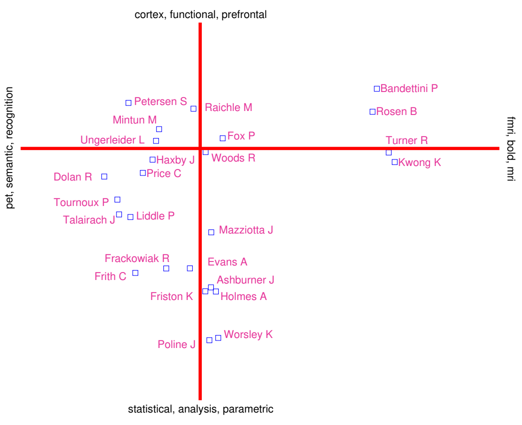
$P = 11558$ author names were extracted from the references. Some of these referenced the same author, e.g., Richard S. J. Frackowiak appeared in 10 misspelled versions. These errors were not corrected as the first few SVD components should not be largely affected by it.



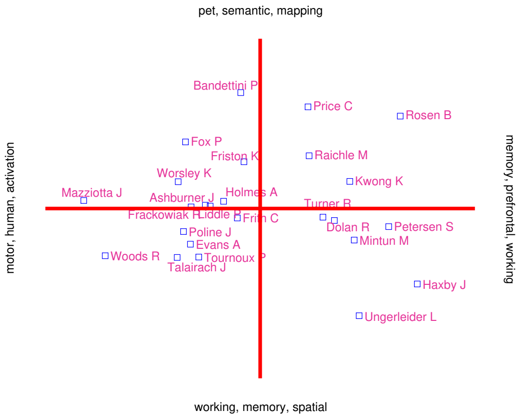
The figure shows authors in the first and second eigenauthor space $\mathbf{v}_1$ (horizontal axis) and $\mathbf{v}_2$ (vertical axis). For clarity only the 25 most cited authors are plotted. Karl J. Friston and Richard S. J. Frackowiak are found as the most authoritative authors while key authors in fMRI-acquisition shows up as a clear cluster well separated by the second eigenauthor: Kwong, Turner, Bandettini and Rosen.

Results

The second $\mathbf{v}_2$ and fourth $\mathbf{v}_4$ eigenauthor are shown below.



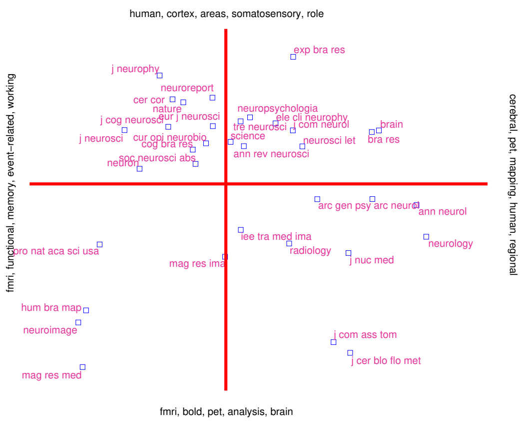
The fourth eigenauthor is associated with the statistical analysis, where Jean-Baptiste Poline and Keith J. Worsley score highest. Others associated with the development of SPM have a high score on this components.



$\mathbf{v}_3$ and $\mathbf{v}_5$ are shown above and the separation is primarily based on the psychological constituent under study: (working) memory in one end and motor activation in the other.

Journal cocitation analysis

Rather than using the cited authors as the basis for analysis the cited journal can also be applied as the basis for an analysis with identifications of *eigenjournals*: An indicator datamatrix $\mathbf{X}(P \times R)$ is constructed from P documents citing R different journals and a similar analysis is performed as with the cited authors. To allow for differences in journal naming/spelling a special stemming is used where journal words are abbreviated to 3 letters, not counting prefixes such as "neuro-" and not abbreviating single word journal names.



The figure shows the most cited journals from *NeuroImage* projected on axes associated with the second (vertical) and third (horizontal) eigenjournals. The axis associated with the second eigenjournal seems primarily to distinguish between macroscopic and microscopic neuroscience with *Magnetic Resonance in Medicine* and *Experimental Brain Research* as the most extreme. The axis associated with the third eigenjournal separates PET and fMRI, — PET being represented primarily with medical speciality journals: neurology, radiology, nuclear medicine and psychiatry. The most authoritative journal based on data from *NeuroImage* is *NeuroImage* itself. The most similar journal is *Human Brain Mapping*.

Remarks, further and future work

- Our method can be automated and performed without human intervention.
- Data from other journals can easily be included though the computational complexity will increase as more data are added.
- Instead of using the title words for labeling of the axes the abstract or (if available) the full text can be used.
- Other forms of analysis can be performed on the data, e.g., ICA, cluster analysis and non-negative matrix factorization (NMF). Below is NMF analysis [3] showing a fraction of authors in 4 of 6 components:

1	2	3	4
Raichle M	Rosen B	Friston K	Haxby J
Petersen S	Turner R	Frackowiak R	Ungerleider L
Buckner R	Bandettini P	Frith C	Mintun M
Grasby P	Kwong K	Evans A	Petrides M
Grady C	Glover G	Talairach J	D'Esposito M
Miezin F	Jezzard P	Worsley K	Gore J
Fletcher P	Belliveau J	Tournoux P	Noll D
Horwitz B	Kim S	Mazziotta J	Cohen J
Posner M	Hyde J	Poline J	Goldman-Rakic P
Maisog J	Ugurbil K	Holmes A	Meyer E
Wise R	Kennedy D	Fox P	McCarthy G
Houle S	Weisskoff R	Dolan R	Koeppe R
Fiez J	Cohen M	Liddle P	Aguirre G
Shallice T	Wong E	Ashburner J	Smith E
Alpert N	Josephs O	Price C	Minoshima S
Herscovitch P	Menon R	Heather J	Corbetta M
Kapur S	Ogawa S	Neelin P	Jonides J
Tulving E	Williams S	Watson J	Nobre A
Paulesu E	Brady T	Marrett S	Shulman G
Rapoport S	Dale A	Phelps M	Paus T
McIntosh A	Jesmanowicz A	Carson R	Zarahn E
Rauch S	Cox R	Collins D	Courtney S
Pardo J	Howard R	Zeki S	Keil K
Craik F	Merkle H	Lammertsma A	Awh E
Pietrini P	Tank D	Myers R	Mishkin M

- It should be possible to identify time trends in the data. However, the data represents a short period of time and we did not examine this issue thoroughly.
- It is possible to develop a method where (query) terms are projected back in cited author space ranking authors conditioned on one or more words. We call this *context sensitive author ranking* and it enables identification of authorities given the words, e.g., in an abstract — with possible application for *automatic identification of reviewers*.

More information

Some of the analytical and visualization tools used in the work are implemented in the *Brede Matlab toolbox* available at `http://hendrix.imm.dtu.dk/software/brede/`

References

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[2] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632, 1999.

[3] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

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

[5] Howard D. White. Cocited author retrieval online: An experiment with the social indicators litterature. *Journal of the American Society for Information Science*, 32:16–22, 1981.

[6] Howard D. White and B. C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32:163–171, 1981.

Acknowledgement

NEC Research Institute. Katherine McCain. This project has been funded by The Danish Research Councils through THOR Center for Neuroinformatics and the Human Brain Project P20 MH57180 "Spatial and Temporal Patterns in Functional Neuroimaging". Daniela Balslev.