Content Coverage of PNAS in 1995 and 2001

Ketan Mane & Katy Börner
Indiana University, SLIS
10th Street and Jordan Avenue
Bloomington, IN 47405 USA
Email: {kmane, katy}@ indiana.edu

This paper reports results of determining changes in content coverage of the Proceedings of the National Academy of Science (PNAS). Two time slices namely 1995 and 2001 were selected as sample data sets. Latent Semantic Analysis [5] was applied to determine semantically similar documents. In addition, co-word analysis was used to examine the occurrence of terms used in titles and keywords of articles. The results were visualized using graph layout algorithms available in the graphical visualization software *Pajek* [1].

Data Set

The utilized data set comprises the complete set of papers in the Proceedings of National Academy of Science for the year 1995 and 2001. For each document the author and publisher assigned keywords and controlled vocabulary Mesh terms from Medline were determined. Statistics on the two datasets are given in Table 1. Obviously papers published in 2001 had little time to receive a significant number of citations.

Table 1: Statistics of the PNAS data sets

Features	1995	2001
Number of documents	2505	2708
Number of unique keywords	14408	15269
Average number of keywords	5.8	5.6
Maximum number of times cited	1447	155

In order to make the data set more manageable, the 500 most cited articles were selected for the subsequent analysis of content coverage. This results in a data set for 1995 in which each document was at least 90 times cited and in the 2001 data set each document received at least 10 citations.

Data Analysis

Latent semantic analysis (LSA), also called latent semantic indexing was applied to determine sets of semantically similar documents [5]. LSA extends the vector space model by modeling term-document relationships using a reduced approximation for the column and row space computed by the singular value decomposition of the term by document matrix [3]. By considering the context of the words, LSA overcomes two fundamental problems faced by traditional lexical matching schemes: synonymy (similar meaning words – impacts recall) and polysemy (words with multiple meaning – impacts precision).

Using code for data parsing and similarity matrix computation available in the information-visualization repository[1] at Indiana University the list of unique terms and the term-by-document frequency matrix were determined for the 1995 and 2001 data set. Subsequently, the LSA SVDPACKC provided by M. Berry [2] was applied to determine the most important latent dimensions. The document-by-document similarity matrices were computed based on the 118 most important dimensions for 1995 and 136 dimensions for 2001.

The word co-occurrence space was created based on the original list of keywords as well as words occurring in the titles of publications.[2] Stop words were eliminated as well as words that occurred less than ten times. The word-by-word similarity matrix was generated based on the co-occurrence of (key)word in the 500 most cited documents for each year. For example, if PROTEIN and GENE are used as keywords in one document then their frequency value is increased by one and hence the similarity increases. The values of the frequency matrices generated for each year were divided by the highest value 271 (1995) and 222 (2001) to obtain the word-by-word similarity matrices.

Data Visualization

A) Semantic Document Space

The Kamada Kawai algorithm [4] implemented in Pajek [1] was used to layout the documents in a 2-dimensional space. The results are displayed in Figure 1 and 2 and are discussed subsequently.

The visualization of the 500 most cited documents published in 1995 shows four major clusters. Three clusters have been labeled *Genomics*, *Botany* and *Molecular Biology* after a careful examination of documents in those clusters. Doc233 (original ID is A1995RB80400002) interconnects different areas with the keywords:

+BREAST-CANCER|+CELL-PROLIFERATION| +COLON CANCER|+DIETARY RESTRICTION|+EPIDEMIOLOGIC EVIDENCE|+HEMATOLOGICAL FINDINGS|+HEPATITIS-CVIRUS| +HEPATOCELLULAR-CARCINOMA|+LOW-INCOMEHOUSEHOLDS| +NON-HODGKINS-LYMPHOMA

[1] http://iv.slis.indiana.edu/

[2] In biological research it is common practice that words occurring in the title cannot be used as keywords. Hence titles that accurately describe the content of a paper imply keywords that may not perfectly fit.

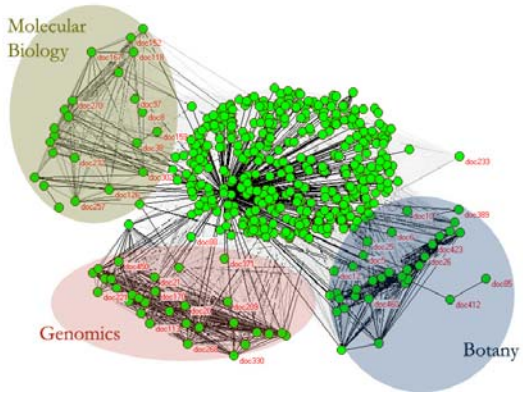

Figure 1: PNAS document space in 1995

The visualization of the 500 most cited papers published in 2001 shows three major clusters. Two of these clusters contain documents describing research in *Genomics* and *Botany*. The more densely packed middle cluster indicates a higher interconnectedness of the papers under study.

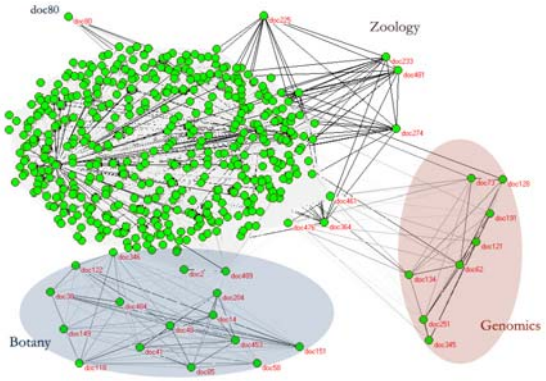

Figure 2: PNAS document space in 2001

Doc 274, 481, 364, 225 (upper right corner) discuss research in *Zoology*. Doc 80 is of special interest as its keywords 'biosynthesis' and 'insect herbivores' bridge between different domains.

B) Co-(Key) Word Space

Figure 3 and 4 show the co-(key)word spaces visualized in *Pajek* using the Fruchterman-Reingold 2D–algorithm [6]. The visualization shows distinct cluster formations of the hot-topics covered in the years under consideration. In consultation with a genomics research expert major clusters have been identified.

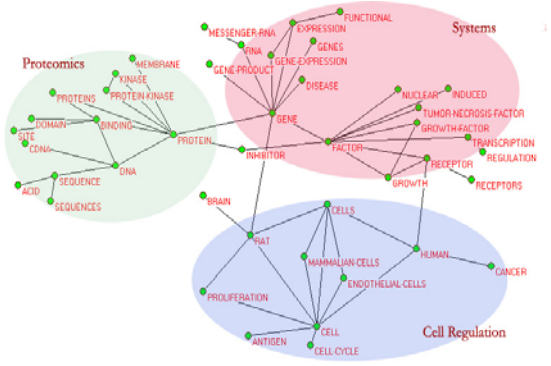

Figure 3: PNAS (key)word space in 1995

As for 1995, displayed in Figure 3, three major clusters exist: *Proteomics* (green), *Cell Regulation* (blue) and *Systems* (red).

The 2001 keyword coverage in Figure 4 shows similar topic coverage but with increasing specialization of research. While the *Systems* cluster in 1995 covered the genomics research, in 2001 *Genomics* (rose) is now a separate filed and more specific clusters like *Gene Regulation* (brown) exist.

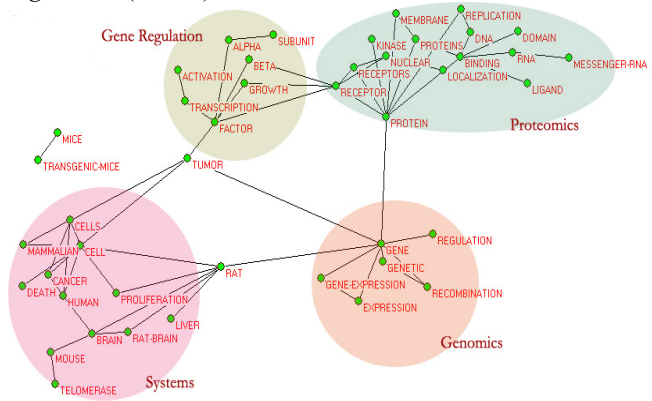

Figure 4: PNAS (key)word space in 2001

Discussion

A visual comparison of the plots leads to the conclusion that there are similar patterns in the usage of keywords in the two years under study. Interestingly, documents appear to form larger clusters in 2001 (large middle cluster) whereas the (key)word space shows an increasing specialization.

A more detailed analysis of the PNAS data examining more than two time slices and involving experts from diverse domains would contribute to uncover detailed patterns of content coverage and content change over time.

Acknowledgements

We would like to thank Anne Prieto for insightful comments on the interpretation of the generated maps.

The data used in this paper was extracted from Science Citation Index Expanded – the Institute for Scientific Information®, Inc. (ISI®), Philadelphia, Pennsylvania, USA: © Copyright Institute for Scientific Information®, Inc. (ISI®). All rights reserved. No portion of this data set may be reproduced or transmitted in any form or by any means without prior written permission of the publisher.

References

1. Batagelj, V. and Mrvar, A. Pajek: Program Package for Large Network Analysis, University of Ljubljana, Slovenia, http://vlado.fmf.uni-lj.si/pub/networks/pajek/, 1997.
2. Berry, M. (1993) SVDPACKC (Version 1.0) User's Guide, University of Tennessee. Available online at http://www.netlib.org/svdpack/
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41* (6), 391-407.
4. Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31* (1), 7-15.
5. Landauer, T.K., Foltz, P.W. and Laham, D. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.
6. Thomas M. J. Fruchterman and Edward M Reingold. (1991). Graph drawing by force directed placement. *Software: Practice and Experience*, 21 (11).