A Comparison of Mapping Algorithms for Author Co-Citation Data Analysis

Haozhen Zhao
College of Information Science and Technology, Drexel University
3141 Chestnut Street, Philadelphia, PA19104
hz58@drexel.edu

Xia Lin
College of Information Science and Technology, Drexel University
3141 Chestnut Street, Philadelphia, PA19104
linx@drexel.edu

ABSTRACT
A key process of any citation analysis study is to map the coded citation data from a high-dimensional dataset to a lower dimensional one while detecting the groups, clusters, patterns or other features of the citation relationships. Over the years, many methods have been used in various studies, including multi-dimensional scaling, Pathfinder networks, Kohonen's self-organizing mapping, etc. Many of these methods are fundamentally different, but their results are similar and comparable. In this study, we selected and applied four of the mapping methods to the same dataset, the author co-citation matrix of the top 100 highly cited information scientists. The results of the different mapping methods provide interesting comparisons among the different mapping algorithms as well as the different views of the dataset.

Keywords
Author Cocitation Analysis, Visualization, Pathfinder Networks, Kohonon Map.

INTRODUCTION
Author Cocitation Analysis (ACA) uses the oeuvres of authors as units of analysis and derives meaningful connections among authors based on the frequencies of their works being cited together. Since its inception, ACA is used in mapping the intellectual structure of a discipline or field (White & Griffith, 1981; White & McCain, 1998) creating visual information retrieval interface for author retrieval (Lin, White, & Buzydlowski, 2003). Over the years, though many methods have been tried in this line of analysis, thorough comparison across different methods is rare. The purpose and intended contribution of this study are to explore possible ways to compare different mapping methods for author co-citation analysis.

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

DATA, METHODS AND RESULTS
Dataset used in this paper is a 100 by 100 author co-citation matrix, of which the rows and columns are the top 100 highly cited authors in Library and Information Science (LIS) during the 1999 to 2008 period. The list is generated by Dr. Katherine McCain following the same procedures in White & McCain (1998). Based on this 100 authors set, we queried the Social SciSearch database in Dialog to obtain co-citation data for the 100*99/2=4950 distinct authors pairs.

The Mapping Process
Mapping ACA data is a problem of two parts. First, the high dimensional square co-citation matrix needs to be reduced to a low dimensional one such that each author will occupy a position in a 2-D or 3-D space. Second, the closeness between author pairs should be preserved at best in the resultant visual maps, showing some kind of grouping and memberships of authors. We applied here four algorithms in the mapping process of our dataset: (1) Multidimensional Scaling with Agglomerative Hierarchical Clustering; (2) Pathfinder Networks; (3) Kohonen Map; and (4) Blondel Community Detection Algorithm.

Multidimensional Scaling with Agglomerative Hierarchical Clustering
In this method, multidimensional scaling is used for ordination and agglomerative hierarchical clustering for grouping authors. We use Pearson *r* as the measure of similarity between authors. The 100 by 100 co-citation matrix is converted to Pearson *r* correlation matrix, before being submitted to multidimensional scaling and agglomerative hierarchical clustering procedures. The procedures for this method are implemented in R, the statistical package.

Figure 1 shows that there are four distinct clusters identified. We label them as Bibliometrics I, represented by ROUSSEAU R, EGGHE L etc., Bibliometrics II (citation) by WHITE HD, MCCAIN KW etc., Information Retrieval by SALTON G, JONES KS, etc., and User Study by BELKIN NJ, BATES MJ, etc. The visualization result of Multidimensional Scaling shows well grouped specialties and two distinct LIS camps, namely the citationists and the retrievalist.

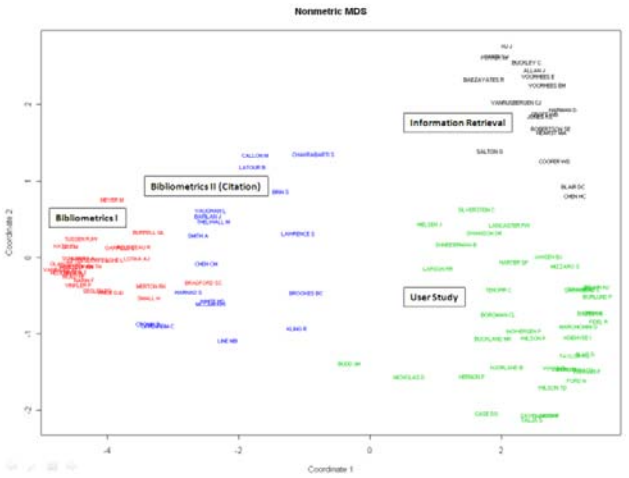

Figure 1. MDS of AHC result

Pathfinder Networks

Pathfinder Networks algorithm approaches the ACA mapping problem as a graph pruning problem. With nodes representing authors, weighted links representing their co-citation counts, the goal is to discard insignificant links while preserving the salient semantic connection patterns in the original network (Schvaneveldt, 1990). Raw co-citation matrix is used in Pathfinder Network algorithm. The result (Figure 2) shows that there are three major clusters identified, with GARFIELD E centered the Bibliometrics cluster, SALTON G the Information Retrieval cluster, and BELKIN NJ the Information Behavior cluster.

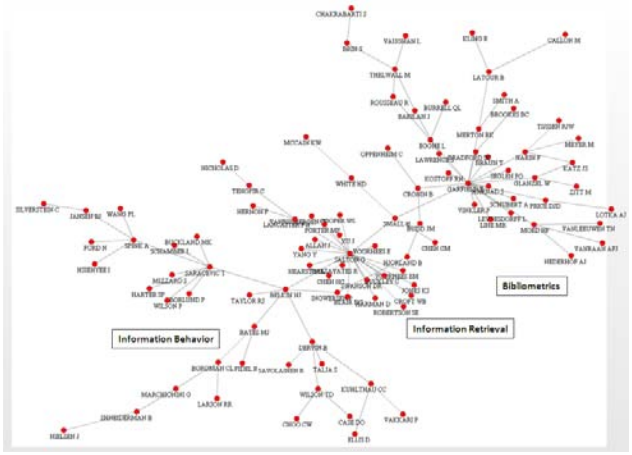

Figure 2. PFNET result

Kohonen Map

Kohonen Map algorithm is an unsupervised learning algorithm in the family of artificial neural networks (Kohonen, 2000). It learns the underlying structure of the original high dimensional inputs in a recursive process and presents the results as rectangle regions. Figure 3 shows our Kohonen Map for the 100 authors. Several distinct regions are labeled, including User Study represented by BATES MJ, KUHLTHAU CC, etc., Information Retrieval by SALTON G, CROFT WB, etc., and Bibliometrics by GARFIELD E, SMALL H, etc.. An interesting group shown explicitly on this map is the Theorist, including WILSON P, BUCKLAND MK, BUDD JM, etc.

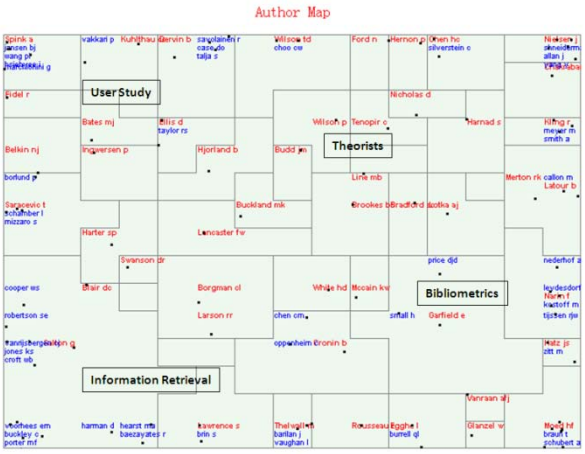

Figure 3. Kohonen Map Result

Blondel Community Detection Algorithm

Community Detection Methods treat the mapping problem as a graph division problem (Newman & Girvan, 2004). We apply on the 100 by 100 co-citation matrix the Blondel community detection algorithm introduced in Wallace, Gingras, & Duhon (2009). Implementation of this algorithm is based on the Network Workbench (NWB Team, 2006). Five communities of different sizes are identified. A visualization using the Circular Hierarchy layout is showed in Figure 4. In additional to the three major clusters, Information Retrieval, Bibliometrics and User Study, which are identified in other clustering methods, another two distinct clusters, Human Computer Interaction and Social Informatics are detected using this method.

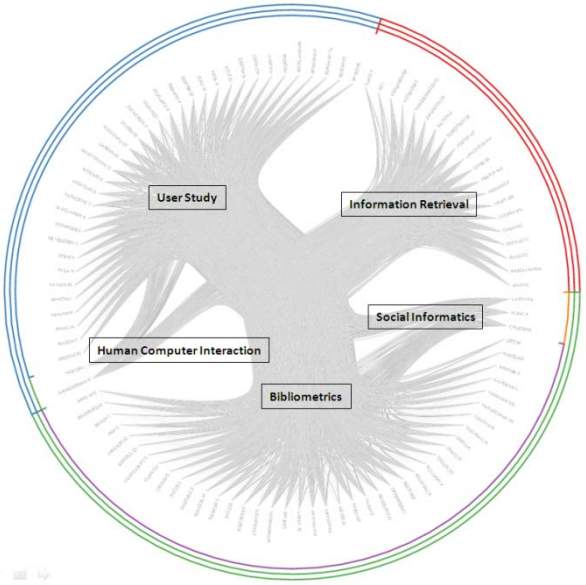

Figure 4 Blondel Community Detection results

DISCUSSION

Comparing our experimental results, we can see that the major specialties in LIS, namely IR, Bibliometrics, and User Study are consistently detected by different methods, implying that ACA is a valid methodology in identifying specialties in the LIS discipline. Beyond this, different algorithms reveal the structure of LIS in different manners: MDS with AHC and Blondel Community Detection give clear global division of the field, while PFNET and Kohonen Map preserve much finer granularity descriptions in terms of the relative positioning of LIS authors.

As for comparison of the different algorithms, we propose the following aspects:

Visual comprehensiveness Among all the mapping layouts, PFNET and MDS are most easily to comprehend, because grouping and membership information can be easily derived from their layout. While Kohonen Maps present richer information about local proximity among authors, it fails to show membership information at a larger scale. For the community detection algorithm, because it does not do any edge pruning, it generates cluttered mapping result.

Interactivity MDS and Community Detection methods need a two-stage processing; one for dimensional reduction, the other for visual ordination, while Pathfinder Networks and Kohonen Map are single-pass processing. Therefore the latter two provide better interactivity for the end users.

Currently, labeling of the generated maps is primarily based on the authors' personal understanding of the field. In-depth comparison among different methods can be enhanced through introducing experts' judgment and providing a more consistent visualization framework.

ACKNOWLEDGEMENTS

We want to thank Dr. Katherine McCain for providing us the top 100 highly cited LIS authors list, and the two anonymous reviewers for constructive suggestions.

REFERENCES

Kohonen, T. (2000). *Self-Organizing Maps* (3rd ed.). Springer.

Lin, X., White, H. D., & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *Information Processing and Management*, 39(5), 689-706.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.

NWB Team. (2006). Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan, http://nwb.slis.indiana.edu

Schvaneveldt, R. W. (1990). Properties of Pathfinder Networks. In *Pathfinder Associative Networks: Studies in Knowledge Organization*. Ablex Pub.

Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology*, 60(2).

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4), 327-355.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-171. doi:10.1002/asi.4630320302