Author Co-citation Analysis is to Intellectual Structure as Web Colink Analysis is to......?

Alesia Zuccala

a.zuccala@wlv.ac.uk
School of Computing and Information Technology, University of Wolverhampton,
Wulfruna Street, Wolverhampton, WV1 1EQ, United Kingdom.

Introduction

Author Co-citation Analysis (ACA) and Web Colink Analysis (WCA) are examined as "sister" techniques in the related fields of bibliometrics and webometrics. Comparisons are made between their data retrieval, mapping and interpretation procedures, focusing on the subject of mathematics. Although the practice of ACA can inform a WCA, the two techniques do not share all research elements in common. The main departure between ACA and WCA exists at the interpretive stage when ACA maps become meaningful in light of citation theory, and WCA maps require interpretation based on hyperlink theory.

Author Co-citations Versus Web Colinks

Author Co-citation Analysis, or ACA, is a specific form of co-citation analysis utilizing highly cocited pairs of *oeuvres*, or selected writings by a sole author or first author in collaboration. ACA was first introduced by White and Griffith (1981) and described in technical detail by McCain (1990). Subsequent authors, notably Persson (2001), Ahlgren, Jarneving, and Rousseau (2003), White (2003), Rousseau and Zuccala (2004) have examined and debated the practice of ACA and offered suggestions for addressing its methodological problems.

In past years, Author Co-citation Analyses have appeared frequently. Scholars are invested in using this technique, yet given the debate current data retrieval/manipulation debate; there has been little or no disagreement regarding interpretations. Lievrouw (1990) reminds us that "author maps reveal the 'cognitive or intellectual structure of a field" and that "the knowledgeable interpreter may see much to explicate in the fine structure of author points: for example common nationality, temporal conjunctions, teacher-student relationships, collegial and co-author relationships, or common philosophical orientations" (p. 103).

Web Colink Analysis (WCA), in comparison to ACA, is a relatively new technique, based on the same pairing principle as its bibliometric "sister" – the pairing of Web colinks instead of bibliographic citations. It may be called the "sister" technique of ACA, because it occupies a position within a subfield of bibliometrics, known as webometrics (Almind & Ingwersen, 1997). The term *colink* in webometrics defines an instance when two Web pages both have inlinks from a third page (Thelwall, 2004). Data collection for a colink analysis requires the use of search engines like AltaVista; however, research to date has focused primarily on the collection and measurement of web page inlinks and outlinks, or directed link networks (e.g., Bjorneborn, 2004).

Colink studies are just beginning to emerge. Larson (1996), for example, conducted an exploratory analysis of a colinked set of Earth Science related websites. Polanco et al. (2001) also used colinks to create a map of 37 European university websites. Thelwall and Wilkinson's (2004) study of a network of academic web domains tested whether or not indirect connections (colinks) on the web would be stronger indicators of subject similarity than direct links. Contrary to their prediction: "high colink counts did not give a higher probability of subject similarity" (p. 66).

Data Selection and Retrieval

For an ACA highly cited/cocited authors are selected and grouped according to a common, yet diversified subject or problem area. For a WCA, highly linked web pages are selected on the basis of a common theme (e.g., academic web pages).

ACA data retrieval requires access to the Dialog™ citation indexes (e.g., SciSearch). The retrieval process is partially automated using a DialogLink™ module and involves the Boolean pairing of cited authors: *S CA=Author, A? AND CA=Author, B?* Mathematical set theory assists in retrieving complete (all-author) cocited author data, although it can be tedious (Rousseau & Zuccala, 2004). Co-citation data is historical in nature – a reflection of the authors' past work.

To retrieve colink data for a WCA, one must use the AltaVista advanced search window. The procedure involves the Boolean pairing of colinks: *link:www.domain.edu AND link:www.domain.edu.* At present it is a manual retrieval process and should be carried out within a day or two. Colink counts on the web are, unlink co-citation data, "up to the minute" and typically fluctuate within a matter of days or weeks.

Co-citation counts as well as colink counts are assembled in an adjacency matrix for analysis. All possible pairs (co-citations and colinks) can reach a maximum of N(N-1)/2. For both ACA and WCA the data scaling debate is applicable, as well as the matrix diagonal problem (i.e., Do we use Pearson's r or Salton's Cosine? Do we treat the cell diagonals as missing values?)

Mapping, Clustering and Interpretation

Figure 1, below, presents a colink map of 44 International Mathematics Research Institutes on the web. With an ACA map, a core-periphery configuration is expected: intellectually similar authors appear close together, while those that are dissimilar sit apart. With a Web Colink Map (Figure 1) the URL core-periphery arrangement is also expected, but a little more difficult to assess: *Author Co-citation Analysis is to Intellectual Structure, as Web Colink Analysis is to…?*

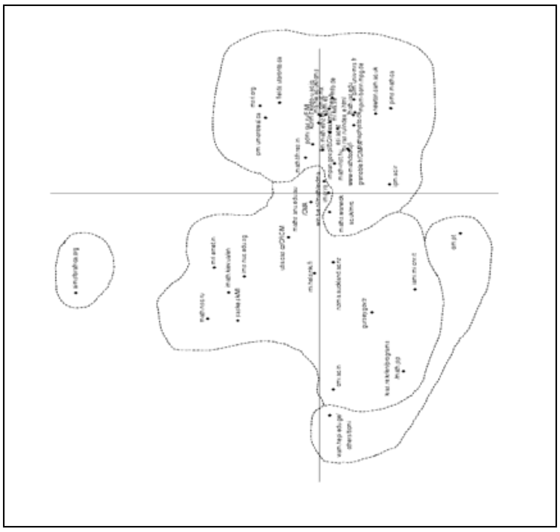

Figure 1. Colink Map of 44 International Mathematics Research Institutes (Nov., 2004).

Link motivation research concerning inlinks to academic websites (i.e., Chu, 2003) has shown that directory-type links are common (i.e., comprising 50%). Since the web page list for the WCA analysis was developed from a Google directory (http://directory.google.com/Top/Science/Math/Research/Institutes/) we expect many directory-based links to exist for similar navigational purposes. Some of the paired institute pages could be producing higher colink counts because they are listed together in more directories. If so, the WCA is generating a map that is somewhat trivial. We might conclude that it is trivial unless we can find another, more significant reason for the colink configuration. The web pages of interest then, are those that have created colinks between the institutes for a purpose other than navigation. To label and interpret the colink clusters, it is necessary to compile lists of colinking pages and examine them for common themes. Are the institute URLs mapped in close proximity colinked more often due to *social* or *personal* reasons than those mapped at distance? To what extent is *geography* playing a significant role? Are some institute pages colinked more frequently due to an underlying *prestige motivation*? Much of the research concerning link theory and motivations for linking is still new; therefore further Web Colink Analyses are needed to understand what makes a web colink structure meaningful.

References

Ahlgren, P., Jarneving, B. & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science & Technology, 54*, 550-560.

Bjorneborn, L. (2003). *Small-world link structures across an academic Web space: A library and information science approach.* Unpublished doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark.

Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50, 65-82.

Chu, H. (2003). *Reasons for sitation (hyperlinking): what do they imply for Webometric research?* Paper presented at the 9th International Con-ference on Scientometrics and Informetrics, 25-29 August 2003, Beijing.

Larson, R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of Cyberspace. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 71-78. Retrieved January 9, 2005, from http://sherlock.berkeley.edu /asis96/asis96.html.

Lievrouw, L. A. (1990). Reconciling structure and process in the study of scholarly communication. In C. L. Borgman (Ed.), Scholarly communication and bibliometrics *(pp. 59-69).Newbury Park, CA: Sage.*

McCain, K. W. (1990). Mapping authors in intellectual space: a technical overview. *Journal of the American society forInformation Science, 41*, 433-443.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics, 50*, 339-344.

Polanco, X. et al. (2001). Clustering and mapping web sites for displaying implicit associations and visualising networks. Patras, Greece: University of Patras. Retrieved January 10, 2005 from http://www.math.upatras.gr/~mboudour/articles/ web_clustering&mapping.pdf

Rousseau, R. & Zuccala, A. (2004). A classification of author co-citations: definitions and search strategies. *Journal of the American Society for Information Science and Technology*, *55*, 513-529.

Thelwall, M. (2004). *Link analysis: An information science approach*. Academic Press.

Thelwall, M. & Wilkinson, D. (2004). Finding similar academic web sites with links, bibliometric couplings and colinks. *Information Processing & Management, 40*, 515-526.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32,* 163-172.

White, H. D. (2003). Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54*, 1250-1259.

.