CAIS Paper: Exploring Author Similarity Using Citing Discipline Analysis

Daisy Jacobs (University of Zululand) and Dietmar Wolfram (University of Wisconsin-Milwaukee)

Abstract: This paper proposes a simple method for assessing author similarity based on the disciplines of citing articles as a complementary approach to more traditional author co-citation analysis. Sixty prolific authors from three allied disciplines are compared using multidimensional scaling and cluster analysis. Distinct and coherent clusters emerge based on disciplines.

Résumé: Cet article propose une méthode simple pour évaluer auteur similitude basée sur les disciplines de citer des articles comme approche complémentaire à plus traditionnelle l'analyse de la co-citation. Soixante auteurs prolifiques de trois disciplines connexes sont comparées à l'aide de positionnement multidimensionnel et regroupement hiérarchique. Groupes distincts et cohérents émergent basé sur les disciplines.

1. Introduction

The study of scholarly communication has been approached from a number of perspectives. Of interest to researchers in scholarly communication and bibliometrics are the relationships that exist between authors based on their publications or how they are used. Author relationships may be studied through patterns of co-authorship (Glänzel & Schubert, 2005), topical similarities in research determined by language use in publications (Lu & Wolfram, 2012), or through citation or co-citation. Citation analysis research has been foundational for studying relationships among bibliographic entities, where the strength of connections between entities is based on the act of citing a work manifested as citations or co-citations (where two entities are cited by a given work). Author co-citation has been studied for decades (see, for example, White & McCain, 1998). Relationships among authors are determined based on the frequency that each pair of authors has been co-cited within publications. Co-citation data can be labour intensive to collect and only reveal relationships between authors that have actually been co-cited. A citation count on its own may reveal something of the impact of a cited author, but it tells us little about the nature of the impact. The origin of a citation, on the other hand, may be revealing of the nature of the contribution or the reach of the cited author. White (2000, 2001) proposed the use of citers to identify characteristics of a given author's research such as an author's citation identity, which consists of all the authors a given author cites. White also introduced the idea of citation image-makers, consisting of the range of authors who refer to a cited author. This list of authors can be thought of as a signature of the cited author, with the frequency of each citing author's citations providing further insight into the influence of the cited author. The signatures of each author may then be compared to identify similarities between cited authors. One challenge with the use of citing authors is that the number can be quite large, which increases the data processing overhead. One way that this overhead may be reduced, while still providing a similar way to compare authors, is to rely on the disciplinary affiliations of the publications of the citing authors instead of the authors themselves using, for example, the Research Area designations for citing journals as determined by a

citation database such as Thomson Reuters Web of Science (WoS). One can reduce the computational burden from potentially thousands of authors to 100-150 disciplinary designations and their frequencies. The feasibility of this approach has been demonstrated in Wang and Wolfram (forthcoming), who examined journal similarity using this approach. This method may also be applied for the comparison of author similarity. The purpose of this study is to explore the applicability of citing discipline analysis for author similarity comparison.

Specific research questions addressed by this exploratory study include:

1) Does the use of visualization methods based on authors' citing discipline signature allow for coherent representation of author relationships?
2) Can interdisciplinary ties between authors in allied disciplines be identified using citing discipline patterns?

This study addresses the conference theme of disciplinary borders. By examining how authors are cited based on the disciplinary origins of the citations they receive, we can better understand the interdisciplinary reach of an author and how authors in the same discipline or allied disciplines may influence one another.

2. Method

Twenty prolific authors each from three allied disciplines included in WoS were identified for study over three time periods. The disciplines were Information Science and Library Science (IL), Communication (CO) and Education & Educational Research (ED). For each discipline, all articles, review articles and conference proceeding articles were identified for the three time periods (1987-1995, 1996-2004, 2005-2012). For each of these disciplines, prolific authors were identified by using the Analyze Results feature of WoS and sorting for number of publications. To minimize issues of data validity arising from of author ambiguity, authors with non-distinctive names were not included in the study.

The data collection method for determining the frequency distribution of citing disciplines used in [Source Withheld] was adopted for the present study. For each author, the "Create Citation Report" option in WoS was selected. The number associated with Citing Articles was then selected to retrieve the list of citing articles. The WoS "Analyze Results" feature was next selected for the list of citing articles. On the Results Analysis page, "Research Areas" were selected as the ranking field to provide the tabulated list of citing disciplines. WoS Research Areas were used to represent the citing article disciplinary affiliation. Citing journals may be assigned to one or more Research Areas. The ranked list of citing disciplines down to a frequency of 2, which represents the most frequently citing discipline signature for each author, was imported into MS Access database management software. In the case of journals with more than one Research Area assignment, each was considered. This simply provides acknowledgement to each area of the these interdisciplinary journals and provides more weight to each discipline. Scripts were written in VBA to calculate the similarity between authors based on the citing disciplines and their frequencies. Salton's Cosine measure was used determine the similarity between pairs of authors, resulting in a symmetric similarity matrix (Ahlgren, Jarneving, Rousseau, 2003).

Multidimensional scaling (MDS) using PROXSCAL analysis and hierarchical cluster analysis using SPSS v.20 were used to visualize and categorize the relationships among the authors for each time period. One advantage of MDS visualizations over some other tools is that it produces a "stress" value, which provides an indication of the goodness-of-fit of the mapping. The lower the stress value, the better the correspondence of the data to the resulting map. Ward's method using squared Euclidean distance was employed to conduct the hierarchical cluster analysis. Outcomes of the cluster analysis were superimposed onto the MDS maps and visually interpreted.

3. Results

For submission length limitations, only the results for the first time period are displayed here. Figure 1 summarizes the MDS and clustering outcome for 1987-1995 with an acceptable goodness-of-fit (Normalized Raw Stress .00817, Stress I .09041). The disciplinary affiliation and number is provided as a label for each author to reduce textual clutter in the figure.
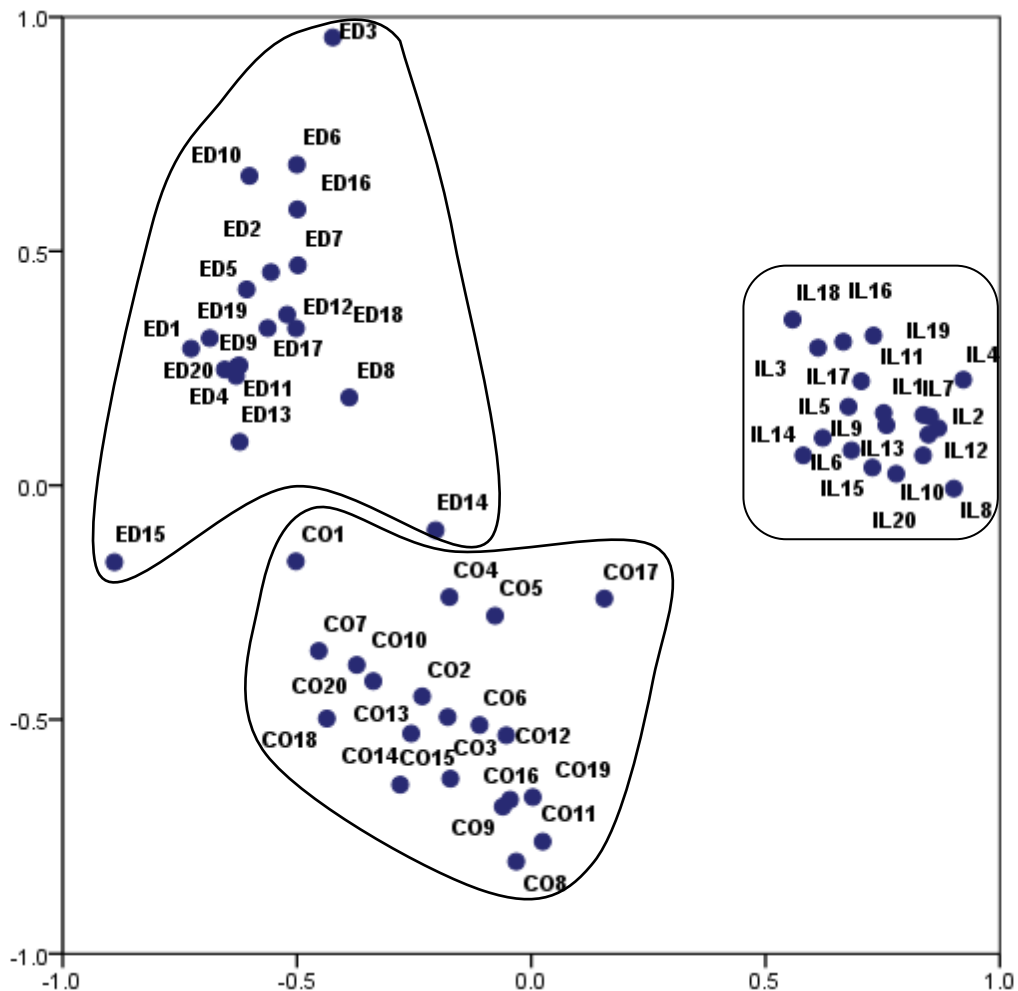

Figure 1. MDS and clustering outcome for authors from three disciplines 1987-1995

One can see from the proximities and the clustering outcomes that prolific authors from each field are clearly more similar to each other than to authors in other disciplines,

except for a small number of cases between CO and ED, where one author affiliated with ED (E14) appears to be more closely aligned with CO based on proximity, but still clusters with ED. A common thread between this author and authors CO4 and CO5 appears to be in the study of perception in each field. In the subsequent time periods, the separations between the clusters representing each field are even more pronounced.

To determine if the citing discipline approach can identify finer-grained distinctions within a discipline, an analysis was also conducted on only the information science and library science authors. Figure 2 summarizes the proximities and cluster assignments for 20 prolific authors in IL for the period 2005-2112. There are four distinct groups revealed by the MDS outcome and corroborated by the cluster analysis, with an acceptable goodness-of-fit (Normalized Raw Stress .00715, Stress I .08458). The clustering outcome reveals a large group of 12 researchers who are known for their work in bibliometrics/scholarly communication, a smaller group of five who investigate information behavior, two researchers who focus on web search, and a single author with an information policy focus. Similar large and small clusterings were observed for the first two time periods, with a large group of researchers broadly categorized under information retrieval and bibliometrics/scholarly communication.
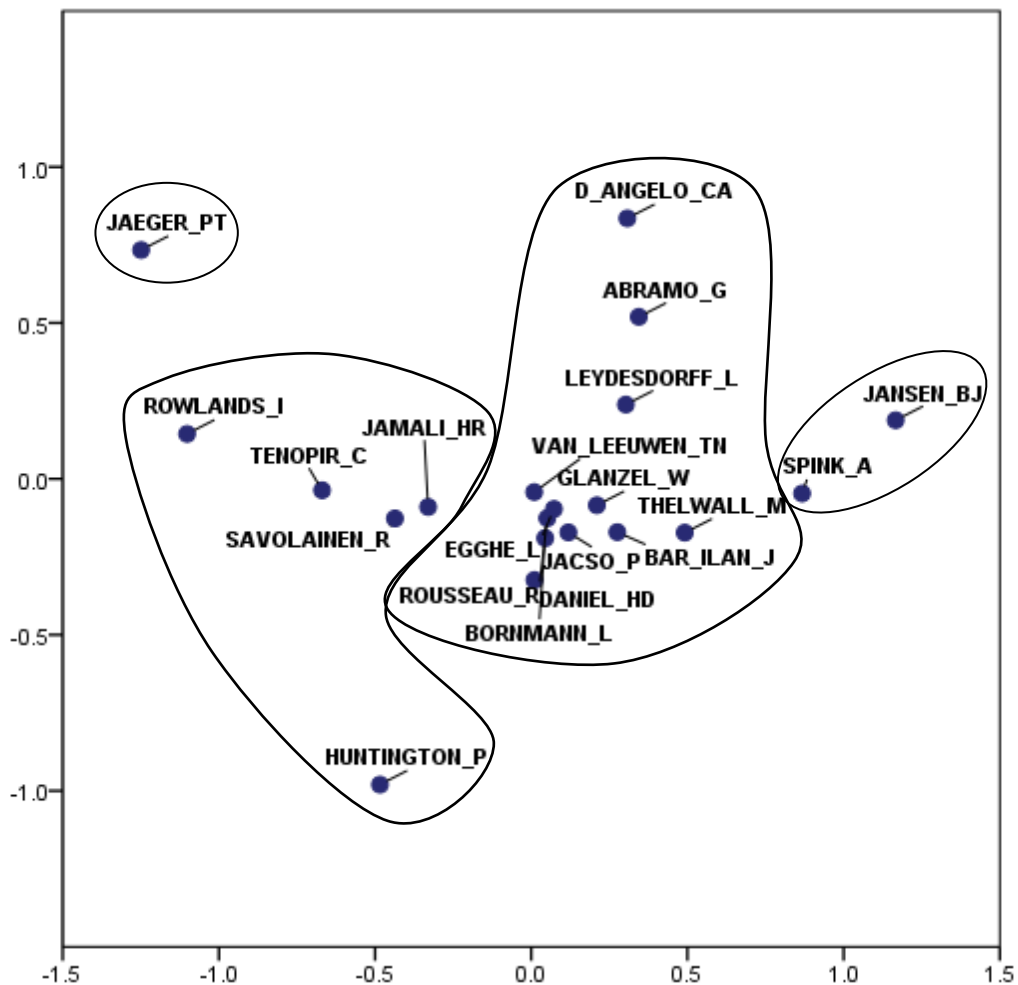

Figure 2. MDS and clustering outcome for prolific information science and
library science researchers 2005-2012

4. Conclusion

Citing discipline analysis offers a complementary way to explore author relatedness to other existing methods such as co-citation analysis. One advantage of the proposed approach is that it does not require authors to be co-cited in order for their similarity to be assessed. However, like co-citation analysis authors must have attracted citations to their work to establish the strength of relationships among the studied authors. Another feature of citing discipline analysis using WoS is that the data collection is less labour-intensive than co-citation analysis (which is not easily supported through WoS) and still provides meaningfully interpretable outcomes. The method may have several limitations. First, the use of disciplinary designations to identify disciplinary relationships may be biased by the human judgments made in assigning journals to one or more given disciplines. As noted in Wang and Wolfram (forthcoming), some journals may not be assigned to the most relevant research areas. Second, the method may not scale well using MDS when comparing disciplines that are not closely related, where authors from allied disciplines may clustered together simply based on relative proximity in comparison to the more distant authors.

Based on the findings of this exploratory study of the prolific authors in the three fields examined, there is little interdisciplinary overlap, where the influence of authors in one field extends to other fields as measured by citations. However, the citing discipline approach, when combined with visualization and clustering methods, also shows clear relationships among the authors studied within information science and library science.

References

Egghe, Leo, & Loet Leydesdorff. 2009. The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the American Society for information Science and Technology 60: 1027-1036.

Glänzel, Wolfgang, & Andreas Schubert. 2005. Analysing scientific networks through co-authorship. In Handbook of quantitative science and technology research (pp. 257-276). Springer Netherlands.

Lu, Kun, & Dietmar Wolfram. 2012. Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. Journal of the American Society for Information Science and Technology 63: 1973-1986.

Wang, Feifei, & Dietmar Wolfram. Forthcoming. Assessment of journal similarity based on citing discipline analysis. Journal of the Association for Information Science and Technology.

White, Howard D. 2000. Toward ego-centered citation analysis. In B. Cronin & Atkins H.B. (Eds.), The Web of Knowledge, (pp. 475–496). Medford, NJ: ASIST Monograph Series.

White, Howard D. 2001. Authors as citers over time. Journal of the American Society for Information Science and Technology 52: 87-108.

White, Howard D., and Katherine W. McCain. 1998. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science 49: 327-355.