Identifying Individual Research Trails

Jochen Gläser[1] and Grit Laudel[2]

[1] *jochen.glaser@fu-berlin.de*
Zentrum Technik und Gesellschaft, Technische Universität Berlin, Hardenbergstr. 36A, 10623 Berlin (Germany)

[2] *g.laudel@rathenau.nl*
Rathenau Institute, P.O. Box 95366, 2509CJ The Hague (The Netherlands)

Abstract

This paper reports the development of bibliometric methods for the analysis of individual research trails. The demand for such methods arose in projects that used qualitative interviews for obtaining data on the content of the interviewee's research. The part of the interview that explores the evolution of the interviewee's "research trail" – the sequence of topics addressed and methods used as it unfolds over time - can be based on bibliometric analyses of the oeuvre of the interviewee. For fields with sufficient ISI coverage, all publications by an interviewee were downloaded and their reference lists automatically searched for shared references and self-citations. In fields with insufficient ISI coverage, title and abstract keywords were identified and compared. On the basis of this information, networks were drawn manually by adding links according to decreasing strength until most publications were linked but clusters could still be distinguished.
The resulting networks fulfilled their role as stimuli for discussions about interviewees' research biographies. Although the networks did not need to be correct in order to stimulate responses, many of them were. Together with the interview process, the networks enabled the identification of actual and latent research trails. Currently, the analysis of oeuvres with latent semantic analysis is tested. The resulting networks will be compared to the results of bibliographic coupling and to the information obtained from interviews.

Introduction

This paper reports the development of bibliometric methods for the analysis of individual research trails. The demand for such methods arose in projects that used qualitative interviews for obtaining information about the content of the interviewees' research. Answering the research questions of those projects required exploring the evolution of the interviewee's "research trail" – the sequence of topics addressed and methods used as it unfolds over time. In the part of the interview that addresses the content of research, the interviewer usually needs to establish the interviewee's "research trail" (Chubin and Connolly 1982), i.e. the history of topics the interviewee has worked on. This is often difficult for interviewers because their knowledge of the interviewee's science is limited at best. When facing this problem, we decided to use structural analyses by bibliometric methods in order to obtain additional information about the interviewee's work We used bibliometric to identify research trails in the oeuvre of the interviewee. These research trails and their visualisations served as points of reference and narration stimuli in the interviews.

State of the art

The literature on qualitative interviewing does not address the use of visual aids or other documents in the interview. The general methodological stance appears to be that whatever generates the necessary responses without orienting or constraining the interviewee is admissible (Laudel and Gläser 2007). The danger of orienting or constraining the interviewee is very low if the interviewees are academics who are used to recognising and countermanding such influences in academic debates.
Bibliometric analyses at the level of individual researchers are limited to a few questions and methods. Evaluative bibliometrics usually regards the individual level as unsuitable because the numbers of publications are too low for the indicators to be valid (Moed and van Raan

1988: 188-190). Analyses of individual oeuvres have so far targeted collaboration (by analysing co-authorships), citing behaviour (White 2000; 2001), visibility (Cole and Cole 1968; Laudel and Gläser 2001), or organisational mobility (Laudel 2003). Individual-level analyses of research content have addressed the degree of interdisciplinarity of papers or projects, mainly by categorising references of papers (Rafols 2007; Rafols and Meyer 2007). Unfortunately, even though bibliometrics has developed a range of methods for analysing epistemic structures of the knowledge represented in publication collections (in particular bibliographic coupling, co-citation analysis, and network and clustering approaches using keywords), these methods are not usually applied by neighbouring science studies disciplines in their investigations of the work of individual researchers.

Approach

Following Chubin and Connolly (1982), 'research trails' can be defined as sequences of thematically interconnected projects in which findings of earlier projects serve as input in later projects. From this follows that researchers' 'portfolios' (the array of actual and latent research trails they can follow) can comprise either one or several distinct parallel research trails. The identification of research trails requires identifying a researcher's oeuvre as either representing one sequence of topically connected projects, or as consisting of several sequences of projects. This task can be operationalised as finding sequences of publications whose internal bibliometric links are stronger than links to other publications in the researcher's oeuvre.

In this paper, we report experiments with bibliometric methods for identifying research trails. Bibliometric analyses of research trails have been combined with interviews in two projects (on research funding in universities and on academic careers). The disciplines covered by the projects include Mathematics, Physics, Biochemistry, Geology, Political Science and History.

For fields with sufficient ISI coverage, all publications by an interviewee were downloaded. The accuracy of the publication list was checked by using data available from the internet (publication lists, organisational affiliations and other biographical information, areas of research). Downloads were processed in EXCEL and automatically analysed for numbers of citations, shared references and self-citations. Ambiguous references were listed separately and inspected manually.

Publications were ordered chronologically and the networks produced by the two types of links were analysed.

1) Networks resulting from self-citations consisted of links of equal strengths but of nodes of varying in-degrees and out-degrees.

2) Because of the low numbers of publications involved in many cases, networks were constructed manually. Nodes representing publications were linked when the publications shared references, beginning with the highest number of shared references. The procedure was ended when most of the publications were linked and introducing links of lower strength led to the unification of clusters in the network. This admittedly imprecise procedure was necessary because the strengths of links between publications as well as the structure of the networks depend not only on field-specific publication and referencing practices but on individual referencing styles as well (White 2001).

Results of 1) and 2) have been compared and turned out to be very similar. Since there were few differences between the networks resulting from self-citations and from bibliographic coupling, and since the latter has the additional advantage of providing information about the relative strength of links, it was used as the method to prepare the interviews.

3) In fields were ISI coverage could be assumed to insufficiently cover the oeuvre of academics (Geology, Political Science, History), publication lists were obtained from academics' web sites and checked for completeness with additional internet searches.

Networks of publications were created by comparing keywords obtained from titles and abstracts of publications.

4) Publication networks resulting from 2 and 3 were visualised by manually drawing a chronological picture (see diagram 1 for an example). Pictures were shown to interviewees in connection with the question about their research biographies since they joined their current university.

Preliminary results

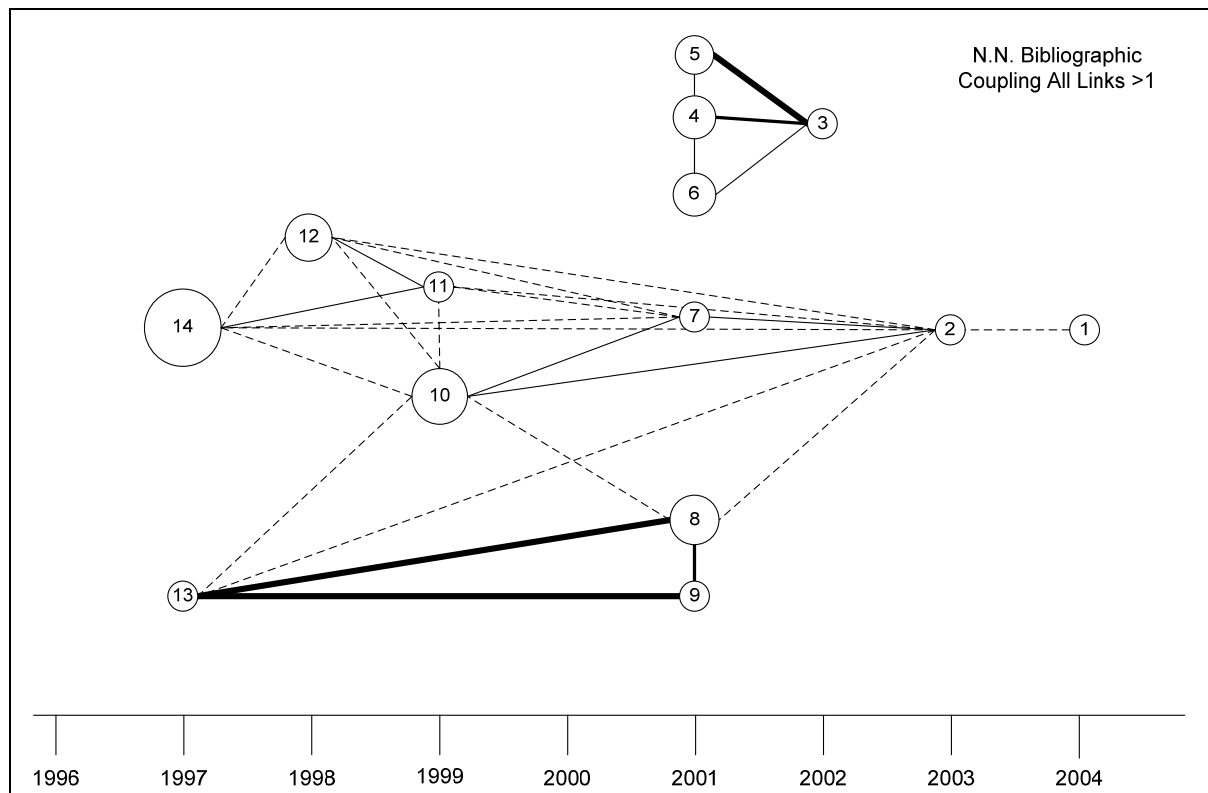Diagram 1 shows a typical picture obtained by bibliographic coupling.

Diagram 1 Example of a bibliometric research trail (strength of lines indicates topical proximity, size of circles indicates numbers of citations

The constructed networks of publications fulfilled their function of stimulating narrations about research biographies and research trails, as is demonstrated by the following quote from the interview with the physicist whose research trails are depicted in diagram 1:

> Yes, yes. So this stuff here [pointing to publications 3-6] , that's when I - my first postdoctoral position, I went to [city in Europe]. And that was stemming from - it was optical physics, so it was similar to some of the physics I've been doing here but from a different point of view, from a classical point of view, so that's why there's not much reference, and I decided to come back to Australia.

Even though the pictures did not need to be valid representations of research trails in order to fulfil their role as narration stimuli (correcting a wrong representation was basically as good a response as confirming a correct representation), it can be noted that many representations turned out to be largely correct. Most corrections provided by interviewees concerned publications that were isolated in the picture because they were linked by very few or no shared references. In some cases, interviewees also stated that research trails that were separated in the picture were in fact not separate, or that something pictured as one research

trail in fact consisted of two different research trail. Corrections of this type were submitted for both ISI-based networks (2) and networks constructed by internet-based analyses (4).

The use of the network visualisations in the interview enabled the identification of 'active' 'latent' and 'abandoned' research trails. Active research trails are pursued at the time of the interview and could be identified because interviewees linked their current projects to research trails depicted by the networks. Latent research trails included those which interviewees had to put on hold against their wishes and wanted to re-activate in the future if opportunities allow (which showed in the networks as clusters of publications that don't reach the present) and those that were planned but not yet started (which interviewees talked about as not yet visible in the network). Abandoned research trails were those that showed in the networks as clusters that don't reach the present and were described by interviewees as something they will not return to.

The results of the bibliometrics-based parts of the interviews indicate that some researchers pursue several research trails in parallel, while others have only one active research trail at a time. The number of research trails pursued by a researcher appears to depend on field, career stage, and performance level of researchers. The relationship between these factors is rather complex. In our sample, high and average performers in biology and physics and average performers in history had more than one distinct research trail, while political scientists, mathematicians, geologists high performers in history and low performers in biology and physics pursued only one research trail at a time.

Apart from the immediate function as narration stimuli, the pictures and accompanying lists of publications also contributed to the atmosphere of the interview. Most interviewees appreciated the effort that went into the preparation of the interview by constructing the networks, and enjoyed the structural perspective on their work provided by the pictures. Several interviewees asked for the pictures and kept them after the interview.

Current work

Bibliographical coupling and key-word based networks have been successfully applied in two projects but have also proven to be not successful in all cases. A major deficit of our current approach is that it includes too many arbitrary decisions that are based on visual inspections of networks. Since correctly identifying research trails is an important task in investigations whose dependent variables describe the content of research, we are attempting to improve methods and to apply novel methods to that task.

We are currently experimenting with a third method that identifies themes in publications by analysing the matrix describing the bipartite network of publications and their references. A set of articles and their cited sources can be seen as a bipartite network where only links between vertices of different kind are allowed. Co-citation analysis links the cited articles using the citing articles, while bibliographic coupling links the citing articles using the cited articles. A method that equally takes into account both these modes of the bipartite network of articles and sources is latent semantic analysis (LSA) that uses the cited sources as semantic units describing themes of a paper. LSA is based on the singular value decomposition (SVD) of the rectangular affiliation matrix that describes the network. LSA can be used to extract latent themes of a bibliography (Mitesser et al. 2008). Its advantage is that more than one theme can be attributed to one paper, which is more realistic than the traditional disjunctive clustering.

We are applying this method to ISI-based oeuvres of interviewees from previous projects and compare the latent themes constructed by LSA to the results from bibliographic coupling and to the parts of the interviews describing research trails. Following the suggestions by reviewers, we will also apply LSA to term-title matrices and compare the results to those obtained by other methods. These comparisons, the results of which will also be presented at

the conference, enable the matching of latent themes to research trails described in the interview. On this basis, properties of the themes can be analysed, and actual themes be compared to latent themes. The interviews that were informed by the visualised bibliographic-coupling networks usually made researchers identify a low number of distinct research trails (between one and five). Single value decomposition returns a larger number of latent themes, many of which are small. We hope that the comparison of the methods (including the interviews) will provide a way to identify current research themes in the list of latent themes.

References

Chubin, Daryl E. & Terence Connolly (1982). Research Trails and Science Policies. Norbert Elias, Herminio Martins and Richard Whitley (Hrsg.), *Scientific Establishments and Hierarchies*. Dordrecht: Reidel, 293-311.

Cole, Stephen & Jonathan R. Cole (1968). Visibility and the Structural Bases of Awareness of Scientific Research. *American Sociological Review* 33: 397-412.

Laudel, Grit, 2003. Studying the brain drain: Can bibliometric methods help? *Scientometrics* 57: 215-237.

Laudel, Grit & Jochen Gläser (2001). Outsiders, Peers and Stars: Analysing Scientists' Integration into Scientific communities with Scientometric Indicators. Frank Havemann, Roland Wagner-Döbler and Hildrun Kretschmer (Hrsg.), *Second Berlin Workshop on Scientometrics and Informetrics 'Collaboration in Science and Technology'*. Free University Berlin, Germany, September 1-3, 2000: Berlin: Gesellschaft für Wissenschaftsforschung, 129-141.

Laudel, Grit & Jochen Gläser (2007). Interviewing Scientists. *Science, Technology & Innovation Studies* 3.

Mitesser, Oliver, Michael Heinz, Frank Havemann & Jochen Gläser (2008): Measuring Diversity of Research by Extracting Latent Themes from Bipartite Networks of Papers and References. In: Hildrun Kretschmar und Frank Havemann (eds.), *Proceedings of WIS 2008,* Berlin.

Moed, H.F. and A. F. J. Van Raan (1988). Indicators of research performance: applications in university research policy. A. J. F. Van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology*. North-Holland: Elsevier, 177-192.

Rafols, Ismael (2007). *Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience.* Vortrag auf der 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, Spain.

Rafols, Ismael & Martin Meyer (2007). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics* 70: 633-650.

White, Howard D. (2000). Toward Ego-Centered Citation Analysis. Blaise Cronin and Helen Barsky Atkins (Hrsg.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Medford, New Jersey: Information Today, Inc, 475-496.

White, Howard D. (2001). Authors as Citers over Time. *Journal of the American Society for Information Science and Technology* 52: 87-108.