Noname manuscript No.
(will be inserted by the editor)

Indirect Bibliometrics by Hypernetwork Analysis

First Author · Second Author

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords First keyword · Second keyword · More

1 Introduction

Semantic analysis does not contain all the information on disciplinary compartmentation nor on patterns of propagation of scientific knowledge as the ones contained in citation networks for example. Furthermore, data collection in the previous algorithm is subject to convergence towards self-consistent themes because of the proper structure of the method. It may give more insight about scientific social patterns of ontological choices in modeling to study communities in broader networks, that would more correspond to disciplines (or sub-disciplines depending on granularity level).

Previous works in quantitative epistemology using various types of networks have shown interesting potentialities. For the citation network, a good predicting power for citation patterns is for example obtained in [?]. Co-authorship networks can also be used for predictive models [?]. A multilayer network approach was recently proposed in [?], using bipartites networks of papers and scholars, in order to produce measures of interdisciplinarity. Disciplines can be stratified into layers to reveal communities between them and

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

Fig. 1 Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection.

therein collaboration patterns [?]. Keyword networks are used in other fields such as economics of technology [?,?].

We describe here a study implementing these ideas for the particular case of a scientific journal for which bibliographical data is difficult to obtain, that is `cybergeo`, an electronic journal in theoretical and quantitative geography, that is concerned with open science issues such as peer-review ethics transparency [?]. Our approach combine semantic communities analysis (as done in [?] but with keyword extraction ; [?] analyses semantic networks of political debates) with citation network to extract e.g. interdisciplinarity measures.

2 Material and Methods

2.0.1 Implementation

The general architecture for data collection is presented in Fig. ??. Citation data is collected from `Google Scholar`, that is the only source for incoming citations [?] in our case as the journal is not referenced in other databases. We are aware of the possible biaises using this single source [?][1], but these critics are more directed towards search results than citation counts.

Text processing is done the same way as in previous section, expect that a particular treatment is done to language detection using *stop-words* and a specific tagger `TreeTagger` is used for other languages than english [?].

3 Results

We show in figures 2 and 3 preliminary results on citation and semantic network. We are able by the reconstruction of the citation network at depth ±1 from the original 1000 references of the journal to retrieve around $45 \cdot 10^6$ references, on which $2.1 \cdot 10^6$ are retrieved with abstract text allowing semantic analysis. We retrieve by community detection in the semantic network typical geographical disciplines, such as :

- Hydrology : `water, basin, river, capac`
- Traffic : `traffic, road, vehicl`
- Biogeography : `habitat, soil, veget, ecosystem`
- Political Science : `polit, cultur, societi, debat`
- Economy : `market, economi, privat, competit, industri`
- Transportation : `transport, travel`
- Teledetection : `cluster, imag, classif, satellit`

[1] or see http:iscpif.frblog/2016/02/the-strange-arithmetic-of-google-scholars/

Fig. 2 Properties of the citation network. Top : Rank-size plot of in-degrees ; three superposing successive regimes must correspond to different literature types or practices across disciplines. Bottom : example of a maximal clique in the citation network, paper of `cybergeo` being in blue.

Fig. 3 Semantic network of concepts in quantitative geography. Corpus consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal `cybergeo` in the citation network. Relevance of keywords were estimated with a bootstrap method, and semantic network is constructed by co-occurrences of keywords (cut at larger degrees, 10% here to delete hubs such as `model` or `space` and efficiently reveal communities).

- Education : `educ, age, student, school`
- Health : `diseas, infect`
- GIS : `gi, geograph inform system`
- Social geography : `neighborhood, resid`

Distribution of keywords within reconstructed disciplines provides an article-level interdisciplinarity, and we can construct various measures at the journal level. Combination of citation and semantic layers in the hyper-network provide second order interdisciplinarity measures. The construction of null models for comparison and the collection of currently missing data (journals for other papers) are currently ongoing so these results are not presented here.