Journal of Informetrics 10 (2016) 1037–1051

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Clustering citation histories in the Physical Review

Giovanni Colavizza[a,*], Massimo Franceschet[b]

[a] Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland
[b] Department of Mathematics, Computer Science, and Physics, University of Udine, Italy

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form 4 July 2016
Accepted 9 July 2016

Keywords:
Citation histories
Clustering
Regression analysis
Physical Review

ABSTRACT

We investigate publications through their citation histories – the history events are the citations given to the article by younger publications and the time of the event is the date of publication of the citing article. We propose a methodology, based on spectral clustering, to group citation histories, and the corresponding publications, into communities and apply multinomial logistic regression to provide the revealed communities with semantics in terms of publication features. We study the case of publications from the full Physical Review archive, covering 120 years of physics in all its domains. We discover two clear archetypes of publications – marathoners and sprinters – that deviate from the average middle-of-the-roads behaviour, and discuss some publication features, like age of references and type of publication, that are correlated with the membership of a publication into a certain community.

1. Introduction

In bibliometrics, the number of citations received by a publication is a rough indicator of the impact of the work among its peers. Several more elaborated citation measures have been proposed. All of them, regardless of the complexity of their defining formulas, assign a publication with a single rating, so that a total ranking among a set of publications can be compiled.

In this paper, we take a different perspective: *citation temporalization*, by considering citation histories (Redner, 2004). There are two main approaches to study the citation history of a publication: synchronous and diachronous. The former approach focuses on the distribution of the publication years of cited publications, the latter on the distribution of received citations over time (Nakamoto, 1988). We mainly focus on the latter: instead of the single number of citations received by a publication at a given time, we consider the full citation history of the publication since its origin. More precisely, the *citation history* of publication i is a vector

$$h_{i,*} = (h_{i,1}, h_{i,2}, \ldots, h_{i,m}),$$

where $h_{i,j}$ is the number of citations received by i during period j and $T = (1, 2, \ldots, m)$ is a series of consecutive temporal periods in some time granularity (e.g., months or years), where we assume 1 to be the period of publication of i. Citation histories extend citation counts by adding a temporal dimension, providing a more informative and less immediate indication of the impact of a publication. While citation counts are snapshots of publication impact at a given time, citation histories move publication impact over time and map a publication's ageing process.

* Corresponding author at: EPFL, Digital Humanities Laboratory, Station 14, INN 116, CH-1015 Lausanne, Switzerland
E-mail addresses: giovanni.colavizza@epfl.ch (G. Colavizza), massimo.franceschet@uniud.it (M. Franceschet).

http://dx.doi.org/10.1016/j.joi.2016.07.009
1751-1577/

The study of patterns of ageing of scientific publications has been very active since decades, its main focus being understanding scientific discourse in different fields and times, and the determinants of the success of a publication. As an example, this problem was posed by Garfield (1980) as one of trying to individuate publications subject to delayed recognition or premature discovery. He framed the task in the following steps: understanding (i) what is a typical citation pattern for every scientific field; (ii) what is a deviation from this typical citation pattern; and (iii) what really qualifies as a premature discovery. Naturally, delayed recognition is but one of the citation patterns which deviate from the typical one. The ageing of scientific literature has also been compared more generally to the process of obsolescence of any kind of phenomena (Pollmann, 2000). The average or typical citation history is linked with information diffusion processes, where several effects interact in causing a considerable amount of information items to go unnoticed, others to be considered for a short amount if time and then fade out (causing the attention peak), others still to remain relevant for longer, even indefinitely (causing the long tail). The speed of recognition, if any, is also driven by intrinsic as well as extrinsic factors (cf. e.g. Line & Sandison, 1974). For example, curves similar to archetypal citation histories are to be found in the proportion of re-shares of Facebook photos during the first hours since upload: even identical photos were found to be associated with very different diffusion "histories" (Dow, Adamic, & Friggeri, 2013).

Citation histories are not meant to rank publications in a compilation. Nevertheless, citation histories associated with different publications can be compared in a more involved way with respect to a total ordering relation. In this paper, we use *clustering* techniques to group citation histories, and hence their corresponding publications, into a set of clusters or communities. Each cluster corresponds to a set of publications with similar citation histories. Hence, the total, non-symmetric ordering relation used to rank publications with citation counts is substituted with a symmetric similarity relation that prescribes which publication belongs to which community. Each community can be represented with its average citation history – we call this aggregated history the *citation macrohistory* of the cluster. Different clusters correspond to different citation macrohistories, and the flexibility of hierarchical clustering methods allows us to tune the granularity of clusters and hence to calibrate the degree of difference of the corresponding macrohistories. Furthermore, we identify a set of determinants, that is independent variables such as the number of received citations, the number of references, the age of references, the number of authors, the length of a publication, and the publication year and type. We use these determinants to elucidate the membership of a publication to a given cluster, in order to provide each cluster with semantics in terms of publication characteristics. We apply the described methodology to the full Physical Review archive, containing more than half a million publications spanning all domains of physics during the last 120 years.

The layout of the paper is as follows. We describe the methodology proposed in this work in Section 2. In particular, we define histories and macrohistories in Section 2.1, we describe the clustering methods adopted to group citation histories in Section 2.2, we discuss the many choices of our experimental setting in Section 2.3, and briefly present the Physical Review dataset in Section 2.4. Section 3 contains the main results of the application of the method to the dataset. In particular, Section 3.1 is devoted to the results of clustering and Section 3.2 identifies the determinants for the detected clusters. Section 4 compares the present work with related literature and Section 5 concludes and outlines further directions of research.

2. Methodology

In this section we formally discuss citation histories as a way to recover the temporalization in received citations. We also introduce and motivate the choice of spectral clustering in order to cluster publications according to their citation histories, present our experimental setup and briefly describe the Physical Review dataset, which will be used as a case study.

2.1. Citation histories

The *citation history* of a publication P tracks the citations that P received since its origin (the date of publication). The events composing this special history are the citations given by younger publications Q towards P, the time of the event being the date of publication of the citing article Q. Suppose, for instance, that P is published in year 2011 and now is end of 2015. If P received 5 citations in 2011, 10 citations in 2012, 3 citations in 2013, 2 citations in 2014, and no citations in 2015, then the citation history of P, using a yearly temporal granularity, is the vector $h_P = (5, 10, 3, 2, 0)$. Notice that sum of the citation history vector components corresponds to the total number of citations accrued by P at the present moment (20 in the example). A publication brought out before P has a longer history, while a publication issued after P has a shorter history.

Formally, let i be a publication, $m \geq 1$ be an integer and $T = (1, 2, \ldots, m)$ be a series of consecutive temporal periods in some time granularity (e.g., month or year), where we assume 1 to be the period of publication of i. For every $j \in T$, let $h_{i,j}$ be the (non-negative integer) number of citations received by i during period j. The citation history of publication i over T is the following vector:

$$h_{i,*} = (h_{i,1}, h_{i,2}, \ldots, h_{i,m})$$

In the following, we discuss relevant choices for the definition of a proper citation history. First of all, what is the *minimum length* of a history? And, related to history length, what is the *minimum number of events* (citations) that define a history?