Journal of Informetrics 9 (2015) 59–68

Contents lists available at ScienceDirect

Journal of Informetrics

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

Unsupervised characterization of research institutions with task-force estimation

Tânia F.G.G. Cova [a], Susana Jarmelo [b], Sebastião J. Formosinho [a], J. Sérgio Seixas de Melo [a,b], Alberto A.C.C. Pais [a,*]

[a] Centre of Chemistry – Coimbra, Department of Chemistry, University of Coimbra, 3004-535 Coimbra, Portugal
[b] Centre of Chemistry – Coimbra, Faculty of Sciences and Technology, University of Coimbra, Rua Sílvio Lima – Pólo II, 3030-790 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 22 August 2014
Received in revised form 29 October 2014
Accepted 6 November 2014
Available online 25 November 2014

Keywords:
Task-force
Scientific productivity
Scientific collaboration
Scientific leadership
Research indicators

ABSTRACT

This work aims at establishing the task-force involved in scientific production at the institutional or national level, globally or per area or sub-area of knowledge. In the proposed system, the estimated task-force is further divided into core (permanent members of the institution(s)) and collaborators (more mobile members), and allows normalization of scientific production.

Research groups/institutions/countries of different sizes/scientific areas can, thus, be directly compared and the time evolution of these groups inspected. Results are presented for the characterization of four universities (from Portugal, Sweden and USA) in the 2008–2012 period, for the research area of Chemistry. It is shown that it is possible not only to estimate the task-force, but also to derive new, relevant indicators for the set under analysis. Aspects pertaining to collaboration fluxes are also assessed.

1. Introduction

In recent years, higher education quality assurance systems have sustained a massive growth, particularly in what concerns research evaluation. The implementation of research assessment practices in the more developed countries has been promoted by the limited amount of economic resources. Additionally, the rapid expansion of science and technology has been coupled to an increasing necessity of evaluating scientific productivity in the various disciplines of knowledge. This has made the measure of scientific output of researchers and institutions an important task for the scientific community.

The highly competitive environment in which research institutions fight for top researchers and research funding (Hazelkorn, 2011), along with the economic constraints that countries are suffering, has led to the development of numerous tools for benchmarking and monitoring research performance at both institutional and individual levels (Cova, Pais, & Formosinho, 2013; Garcıa, Rodriguez-Sanchez, Fdez-Valdivia, Robinson-Garcıa, & Torres-Salinas, 2013).

There are different ways to assess science and information flows. The difficulty is to clearly understand what good science is and to know if the chosen data reflects quality (Vanclay & Bornmann, 2012). In this context, the establishment of national and international rankings (Lundberg, 2006) and the introduction of national research assessment exercises in

* Corresponding author. Tel.: +351 239854466; fax: +351 239827703.
E-mail address: pais@qui.uc.pt (A.A.C.C. Pais).

http://dx.doi.org/10.1016/j.joi.2014.11.005
1751-1577/

different countries (Abramo, D'Angelo, & Di Costa, 2011; Lundberg, 2006; Vanclay & Bornmann, 2012) have centered the development of novel methodologies for research monitoring in the focus of political agendas. However, opinions differ on what concerns the concepts, implementation and validity of those rankings.

Often, the research output of certain institutions cannot be accurately determined as a consequence of the existence of some unobservable details or because not all the relevant factors governing the structure of that institution are taken into consideration; nevertheless, many studies can be found in the literature developing techniques to classify research institutions (Ortega, Lopez-Romero, & Fernandez, 2011; Shin, 2009), establish institutional profiles (Carpenter et al., 1988; Garcıa, Rodrıguez-Sanchez, Fdez-Valdivia, Robinson-Garcıa, & Torres-Salinas, 2012) or compare institutions performance (Adams, Gurney, & Marshall, 2007; Tijssen, van Leeuwen, & van Wijk, 2009; Torres-Salinas, Moreno-Torres, Delgado-Lopez-Cozar, & Herrera, 2011).

Research evaluation at the national level is extremely important as a tool for encouraging scientific productivity, particularly if the results are used to support selective funding decisions by government institutions (Sahel, 2011; Turner, 2007). At the same time, it is necessary to avoid conceptual and methodological problems in the assessment of the institutions.

Bibliometrics has become an indispensable tool in the evaluation of institutions and, together with peer or expert judgments, can generate comprehensive and reliable quantitative data for fair assessments (Bornmann & Marx, 2013). It contributes to describe the thematic profile of research units, identify their strengths, analyze their collaboration practices and study trends over time (see e.g., D'Angelo, Giuffrida, & Abramo, 2011; OST, 2010). At the same time, technical limitations inherent in the available bibliometric databases have held back the diffusion of bibliometrics as a "noninvasive" support system for the evaluation of research. These limitations are related to the difficulties involved in correctly identifying the true authors and institutions of each publication, particularly because of homonyms among names, variations in the way individual authors indicate their name and affiliation and incorrect database information. Many studies have reported the need to clean up databases in order to turn the institutional uniformization of names a future reality (Bador & Lafouge, 2005; Mallig, 2010; Morillo, Santabarbara, & Aparicio, 2013).

Methods to disambiguate author names are usually categorized as supervised and unsupervised methods. Supervised methods require manually labeled data to train the algorithm before disambiguating each author instance. The training set can be used to learn characteristics of each author or a generic similarity metric between instances of the same author. These have been recently explored to build a probabilistic model for estimating the probability that a pair of author instances refers to the same individual (Torvik, Weeber, Swanson, & Smalheiser, 2005). The majority of other supervised approaches use training data for each author to disambiguate. This method usually yields better results as the disambiguation algorithm has specific information on each ambiguous author. Another key advantage is the ability to leverage well-established machine learning technologies like classification and clustering (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004). The major drawback of supervised approaches is the need for a training set. This assumption is expensive in practice, and manual labeling of data can become impractical for large scale bibliometric databases. Moreover, maintaining the training set may be prohibitive when data change frequently. To address these issues, some unsupervised approaches have been proposed (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Han et al., 2004; Han, Xu, Zha, & Giles, 2005; Song, Huang, Councill, Li, & Giles, 2007; Wooding, Wilcox-Jay, Lewison, & Grant, 2006). These methods do not need manually labeled data for training the disambiguation algorithm. They formulate the author name disambiguation problem as a clustering task, where each cluster contains all the articles by the same author. In this case, the distance metric is not learned by a training set but it is given directly by the model employed (Han et al., 2005; Song et al., 2007). Other unsupervised approaches have used collaboration or citation graphs to disambiguate authors (McRae-Spencer & Shadbolt, 2006). Some important shortcomings in existing approaches are poor scalability and expandability properties. Most algorithms cannot be used efficiently in large-scale bibliographic databases and cannot properly handle frequent changes to the database (Huang, Ertekin, & Giles, 2006; On, Lee, Kang, & Mitra, 2005). In simple terms, different techniques have been proposed to deal with entity problems, without a simple and direct solution.

A scientific article and its citations (in other articles) represent the increment of new science and the credit for its discovery. Articles and citations are useful measures to assess the productivity of researchers, research groups, research institutions and even countries. The number of citations received by one article is a direct measure of its usefulness to other researchers. As a consequence, new research fields and businesses that seek to develop algorithms to refine articles and citations into a quantifier that reflects scientific productivity or quality, have emerged (D'Angelo et al., 2011; Davis & Papanek, 1984; Egghe, 2006; Garfield, 2006; Hirsch, 2005; Kinney, 2007; Nejati & Jenab, 2010; Moed 2008). However, it is not clear which of the different techniques should be preferred, because they do not, individually, provide a satisfying answer.

Some normalization is needed to compare raw numbers, scientific profiles, institutional groups and research contributors during a defined time period. This requires the definition of strategies to estimate size. Institutional size-based indicators may thus be defined using the most relevant groups included in the task-force of the institutions. This overcomes the problem inherent to the average-based indicators among other measures currently used to analyze and rank research institutions (see e.g., Bornmann, 2013; Bornmann & Mutz, 2011; Bornmann, Mutz, Marx, Schier, & Daniel, 2011; SCImago Reseach Group, 2012; Waltman et al., 2012).

A significant amount of work has been published on collaboration (see e.g., Almeida, Pais, & Formosinho, 2009; Leydesdorff, Wagner, & Adams, 2013), with aspects ranging from cultural and geographical proximity, to coauthorship relations in the ranking schemes of several countries. In this respect, it is suggested in this work to inspect the authors of each article and identify those in which the corresponding author belong to the focused institution (or not), those that