Journal of Informetrics 12 (2018) 133–152

Contents lists available at ScienceDirect

Journal of Informetrics

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

Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics

Peter Sjögårde[a,b,*], Per Ahlgren[c]

[a] Department of ALM, Uppsala University, Uppsala, Sweden
[b] University Library, Karolinska Institutet, Stockholm, Sweden
[c] School of Education and Communication in Engineering Sciences (ECE), KTH Royal Institute of Technology, Sweden

ARTICLE INFO

Article history:
Received 27 September 2017
Received in revised form
15 December 2017
Accepted 15 December 2017

Keywords:
Algorithmic classification
Article-level classification
Classification systems
Granularity level
Topic

ABSTRACT

The purpose of this study is to find a theoretically grounded, practically applicable and useful granularity level of an algorithmically constructed publication-level classification of research publications (ACPLC). The level addressed is the level of research topics. The methodology we propose uses synthesis papers and their reference articles to construct a baseline classification. A dataset of about 31 million publications, and their mutual citations relations, is used to obtain several ACPLCs of different granularity. Each ACPLC is compared to the baseline classification and the best performing ACPLC is identified. The results of two case studies show that the topics of the cases are closely associated with different classes of the identified ACPLC, and that these classes tend to treat only one topic. Further, the class size variation is moderate, and only a small proportion of the publications belong to very small classes. For these reasons, we conclude that the proposed methodology is suitable to determine the topic granularity level of an ACPLC and that the ACPLC identified by this methodology is useful for bibliometric analyses.

1. Introduction

Classifications of scientific publications have multiple purposes. In libraries, publications can be classified and arranged according to a classification scheme to help users browse a physical collection by subject area.[1] Classifications can also be used within libraries to study circulation statistics or downloads. In the digital world, a classification scheme can be used for information retrieval tasks with the purpose to identify relevant documents for a user, e.g. by refining search results to one or more categories in the classification. Within the bibliometric practice at higher education institutions, classification of research publications can be used to study the structure and processes of research activities and to evaluate research in different subject areas.

Traditional classification schemes used in libraries, such as the Dewey Decimal Classification (DDC) or the Universal Decimal Classification (UDC), were created before the digital era. They were created for shelf arrangement and browsing of

* Corresponding author at: University Library, Karolinska Institutet, 17177 Stockholm, Sweden.
E-mail addresses: peter.sjogarde@ki.se, perahl@kth.se (P. Sjögårde).

[1] We use the term "subject area" in a broad sense, to denote an area of research of any level of aggregation. This could be broad areas such as "Computer Science" or more narrow areas such as "Robotic Sensing".

https://doi.org/10.1016/j.joi.2017.12.006
1751-1577/

physical publications. Each publication was classified manually and placed at the corresponding shelf. The classification was documented on library cards which enabled retrieval of publications by subject area. The granularity of the classification, i.e. how finely or coarsely the classification is grained into classes, had to be set in relation to this physical context. Large, specialized library collections had (and still have) a need for finely grained classifications. Small, general library collections had (and still have) a need for more coarsely grained classifications. The commonly used classification schemes meet these diverse demands by their hierarchical structure. Libraries with large, specialized collections can classify publications at a finely grained level while libraries with small, general collections can use the same classification scheme at more aggregated levels.

Historically, the physical research journal was classified into classes using the traditional classification schemes. However, individual research publications were not classified, other than assigning them into the same class as the journal issue in which they had been published. This was a natural consequence of the physical media, because publications were physically bound to a journal issue. Today, research publications are born digital and a large proportion of research publications that were published as physical publications the last decades have been digitized. This transition has opened for new possibilities to analyze bibliographic data, which in turn have led to an increased interest in quantitative studies of research publications. As a response to an increased demand for such studies, the research and professional fields of bibliometrics have grown, in particular the last decade. To be meaningful, bibliometric studies commonly require research publications within different broad fields to be classified into narrower areas, and the granularity of the classification is dependent on the purpose of the study.

In our daily practice as bibliometric analysts at a Swedish university, we have regularly received questions from researchers about, e.g. publication quantities, highly cited papers and/or co-publishing. The questions have often been related to specific subject areas, sometimes broad and sometimes narrow, and not uncommonly both; broad to get a comprehensive picture, and narrow to be able to zoom into more finely grained subject areas.

Until a few years ago, the alternatives for subject classification were few. The traditional classification of journals had not been constructed to meet the demands made by the new data analysis practices. These practices require the classification to be comprehensive, uniformly applied through the data collection and to follow a clearly defined set of rules so that the assignment of publications is not dependent on subjective judgements of the classifier.

Alternatives to the traditional classification schemes are applied in the, nowadays web-based, citation indexes. Citation indexes were proposed by Eugene Garfield in 1955, and Web of Science was developed in the 1950s and 60s (Garfield, 1955, 1964). Parallel to the development of the Journal Citation Reports (JCR), where journals are ranked according to citation rates (Garfield, 1972), journal categorization was created (Pudovkin & Garfield, 2002). The JCR categories were based on similar methods as the classification performed using traditional classification systems, later called a "heuristic procedure" by Pudovkin and Garfield (2002). More advanced approaches have been proposed for journal classification in recent decades. These approaches use citation relations between journals for their classification (Archambault, Caruso, & Beauchesne, 2011; Boyack, Klavans, & Börner, 2005; Chen, 2008; Doreian, 1988; Leydesdorff, 1987, 2006; Leydesdorff, Bornmann, & Wagner, 2017; Pudovkin & Garfield, 2002; Rosvall & Bergstrom, 2011; Small & Koenig, 1977; Zhang, Liu, Janssens, Liang, & Glänzel, 2010).

The many limits of journal-level classification have been acknowledged in the literature (Archambault et al., 2011). An obvious problem is that some journals are broad in scope and thus include publications within different subject areas. Hence, a single subject category cannot accurately represent the subject contents of all publications in such journals. One proposed solution for this problem has been to classify publications appearing in multidisciplinary journals into journal categories created in preceding steps (Glänzel, 2003; Glänzel, Schubert, & Czerwon, 1999; Glänzel, Schubert, Schoepflin, & Czerwon, 1999; Gunnarsson, Fröberg, Jacobsson, & Karlsson, 2011). However, this approach solves the problem only partially. In view of this, publication-level classifications are desirable. Considering the high number of publications, manual approaches to publication-level classifications are time consuming and demand enormous amount of resources. Also algorithmically constructed publication-level classifications of research publications (ACPLCs) require a lot of resources, in this case computational resources, much more than journal level classifications. Until recent years, such classifications have been created merely for small or medium size publication sets.

Global[2] subject maps of science have been shown to be more accurate and useful than local maps (Boyack, 2017; Klavans & Boyack, 2011; Rafols, Porter, & Leydesdorff, 2010). Similarly, global classifications have some of the same advantages. For example, they may be useful for studies (a) where subject differentiation is of importance, (b) dealing with identification and analysis of emerging research fields (Milanez, Noyons, & de Faria, 2016; Small, Boyack, & Klavans, 2014), and (c) aiming to reveal relations between subject areas. Local, small or medium scale mappings or classifications do not provide the same possibilities. To facilitate such studies, global publication-level classifications have been constructed in recent years (Klavans, 2014a, 2014b; ; Šubelj, van Eck, & Waltman, 2016; van Eck, 2012, 2013a;). This development is a huge step forward in the area of research classification. Nevertheless, the methods for ACPLCs are in need for development. In this article, we will address one of the challenges that hitherto have been addressed only briefly.

[2] "Global" refers to a comprehensive coverage of subject areas. Similarly, "local" refers to the coverage of one or a few related subject areas.