Journal of Informetrics 9 (2015) 273–284

Contents lists available at ScienceDirect

Journal of Informetrics

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

Problems and challenges of information resources producers' clustering

Anna Cena[a,*], Marek Gagolewski[a,b], Radko Mesiar[c,d]

[a] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland
[b] Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland
[c] Faculty of Civil Engineering, Department of Mathematics, Slovak University of Technology, 81 368 Bratislava, Slovakia
[d] National Supercomputing Center IT4Innovations, Division University of Ostrava IRAFM, 30.Dubna 22, 701 03 Ostrava 1, Czech Republic

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form 6 February 2015
Accepted 6 February 2015
Available online 25 February 2015

Keywords:
Aggregation
Hierarchical clustering
Distance
Metric

ABSTRACT

Classically, unsupervised machine learning techniques are applied on data sets with fixed number of attributes (variables). However, many problems encountered in the field of informetrics face us with the need to extend these kinds of methods in a way such that they may be computed over a set of nonincreasingly ordered vectors of unequal lengths. Thus, in this paper, some new dissimilarity measures (metrics) are introduced and studied. Owing to that we may use, e.g. hierarchical clustering algorithms in order to determine an input data set's partition consisting of sets of producers that are homogeneous not only with respect to the quality of information resources, but also their quantity.

1. Introduction

Classically, unsupervised machine learning techniques are applied on data sets with fixed number of attributes (variables). Clustering (see e.g. Hastie, Tibshirani, & Friedman, 2009) is among the most popular class of methods of this kind. The main aim of clustering is to automatically discover groups, called clusters, of objects in such a way that entities within each group are similar and objects in distinct groups differ – with respect to some criteria – as much as possible from each other. It is frequently applied in practice, also in informetrics. For example, Nieminen, Pölönen, and Sipola (2013) uses clustering techniques to discover main topics discussed in current data mining research literature. Moreover, Waltman, van Eck, and Noyons (2010) investigated groups in bibliometric networks.

However, many problems encountered in the field of informetrics face us with the need to extend these kinds of methods in a way such that they may be computed over a set of nonincreasingly ordered vectors of unequal lengths. One of the main informetric tasks aims to evaluate the quality of information resources and their producers. In the so-called Producers Assessment Problem (PAP, cf. Cena & Gagolewski, 2013) we assume that we have a set of l producers and that the i-th producer outputs n_i products. Additionally, each product is given some kind of rating concerning its overall quality. Consequently, the state of an i-th producer may be represented by a nonincreasingly ordered sequence of real numbers. Firstly, please note that in PAP lengths of vectors may vary from producer to producer and that the distribution of the producers' productivity (as well as the items' quality) is most often highly skewed. For example, if a scientist is considered as an information resources' producer, then his/her scholarly papers are conceived as products in the PAP model. Here, the papers' quality is

* Corresponding author. Tel.: +48 223810378.
E-mail address: cena@ibspan.waw.pl (A. Cena).

http://dx.doi.org/10.1016/j.joi.2015.02.005
1751-1577/

often described by a function of the number of citations they received (see e.g. Franceschini & Maisano, 2009; Gagolewski & Mesiar, 2012). Similarly, a Facebook or Twitter user is also a kind of producer. In such a case, his/her posts are products, and the numbers of their "re-tweets" or "likes" can be considered as their quality assessment. Secondly, even though one often assumes that products' valuations are nonnegative (or even integer), this is not always the case. For example, Stack Exchange (http://stackexchange.com/) is a network of over one hundred communities created and run by enthusiasts and experts on specific topics like computer science, physics, linguistics, beer tasting, parenting, philosophy, etc. Basically, Stack Exchange contains question and answer sites focused on each community's area of expertise. In each of these sites a user produces posts – questions and answers – which are rated by the community members with not only so-called UpVotes (+1), but also DownVotes (−1). Thus, the assessment of some answer may be negative.

In this paper we discuss clustering algorithms that may be applied on informetric data in order to automatically discover diverse groups of producers. Such methods are crucial not only in the identification and/or description of certain groups of producers (productive, high impact, low impact, etc. ones), but also it may be used in automated informetric decision support systems. One of the possible approaches to apply clustering techniques on vectors of nonconforming lengths is to reduce the data dimension by considering a fixed number of attributes or indicators. For example, Ortega, López-Romero, and Fernández (2011) performed an automatic categorization of universities basing on several indicators. Here, research institutions were grouped according to their outputs described by 13 indicators such as the number of published papers, amount of funds obtained, number of patents, etc. Firstly, principal component analysis (PCA) was applied, and then an agglomerative hierarchical clustering method was used to calculate an a priori categorization of the institutes according to their scores in the obtained components. Similarly, Cheng and Liu (2006) studied a data set on the 500 best world universities in order to divide them into groups according to their various bibliometric performance indicators. Moreover, Costas, van Leeuwen, and Bordons (2010) in order to split a group of scientists into three clusters (top, medium, low class ones) used, e.g. the h-index (Hirsch, 2005), number of publications, number of highly cited papers, median impact factor, etc. Similarly, Ibàñez, Larrañaga, and Bielza (2013) used, i.a. total number of papers and total number of citations for that purpose. Another approach is to pad the input data with some fixed values, so that we get vectors of the same lengths. As most of the clustering techniques are based on distance metrics that are defined by considering the differences between corresponding vectors' elements, the value 0 may be used for that purpose, as $a - 0 = a$ and $0 - b = -b$. In fact, this is what happens when we, e.g. use the h-index: it gives the same values for a given citation sequence and for the same sequence but with trailing zeros.

Unfortunately, the problem with the first approach is that uncountably many impact indices may be used here. For instance, the h-index, the g-index (Egghe, 2006), the w-index (Woeginger, 2008) or their simple modifications may be utilized. The decision on which set of indices should be selected or even what properties such tools should possess is always very difficult (see e.g. Gagolewski, 2013 and Schreiber, 2013). Furthermore, the problem of choosing appropriate number of indices in order to capture *all necessary* information about input data is not easy to solve. What is more, in case of clustering, one often standardizes or scales attributes' values. However, it is known that some impact indices (like the h-index and its generalizations; see Cena & Gagolewski, 2013; Gagolewski & Mesiar, 2012) may be very sensitive with respect to simple input data transformations. Taking the above facts into account, it is clear that a "projection" approach might not lead to plausible results.

On the other hand, the second approach is also quite problematic. By simply padding vectors with zeros, we loose some information. Let us consider a scientist A with one paper cited one time and another scholar B with one paper cited one time and 10 papers with no citations at all (perhaps because he/she is a young researcher). In this case, both authors are indistinguishable: no credit is given to B for being more productive. Of course, ideally we would like to be able to somehow distinguish such vectors from each other.

In this paper we introduce a class of metrics (dissimilarity measures) that can be directly applied to vectors of nonconforming lengths, see Section 2. Owing to that, we are able to retain all the information in the input data. Using such metrics, hierarchical clustering techniques, cf. Section 3, can be directly applied. According to numerical experiments on various empirical data sets performed in Section 4, we find that the new approach distinguishes producers of different impact and productivity well. Moreover, interesting future research directions are addressed in Section 5.

2. Metrics

For any m, let $\mathcal{S}_m$ denote the set of nonincreasingly ordered real vectors of length m, i.e. $\mathcal{S}_m = \{(x_1, \ldots, x_m) \in \mathbb{R}^m, x_1 \geq \ldots \geq x_m\}$. Moreover, let $\mathcal{S}_{\leq m}$ be a set of nonincreasingly ordered vectors of length at most m, that is $\mathcal{S}_{\leq m} = \bigcup_{i=1}^{m} \mathcal{S}_i$.

Assume that we are given l producers and that each of them produced no more than m products for some m. Obviously, such m is finite and well defined for each set of producers. The set of producers may thus be represented by $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(l)}\}$, where $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_{n_i}^{(i)}\right) \in \mathcal{S}_{\leq m}$ for all $i = 1, \ldots, l$. For instance, $x_j^{(i)}$ may denote the number of citations of the j-th most cited paper of the i-th scholar, or the score of the j-th best post by the i-th Stack Exchange user.

In a clustering task, we are interested in finding a partitioning of $\mathcal{X} = C_1 \cup C_2 \cup \ldots \cup C_k$, where $C_1, \ldots, C_k$ are mutually disjoint subsets of $\mathcal{X}$. Most of the clustering techniques are based on computing the degree of dissimilarity between every two vectors in a given set. To measure the dissimilarity, the notion of a metric (distance) is most often used. A *metric* on a set Z is a function $d : Z \times Z \to [0, \infty)$ such that for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in Z$: