Journal of Informetrics 10 (2016) 789–813

Contents lists available at ScienceDirect

Journal of Informetrics

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

Gazing at the skyline for star scientists

A. Sidiropoulos[a], A. Gogoglou[b], D. Katsaros[c,d,*,1], Y. Manolopoulos[e]

[a] Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, Greece
[b] Department of Informatics, Aristotle University of Thessaloniki, Greece
[c] Electrical Engineering Department & Yale Institute for Network Science, Yale University, United States
[d] Electrical and Computer Engineering Department, University of Thessaly, Greece
[e] Department of Informatics, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 8 April 2016
Accepted 10 April 2016

Keywords:
Skyline operator
Perfectionism index
Hirsch index
Scientometrics

ABSTRACT

Admittedly, despite the plethora of scientometric indices proposed to rank scientists, none of them can fully capture the performance and impact of a scientist, since each index quantifies only one or a few aspects of his/her multifarious performance. Therefore, the task of scientometric ranking can be seen as a multi-dimensional ranking problem, where the different indices comprise the dimensions. The application of the skyline operator comes then as a natural solution to the problem. In this article we apply the skyline operator to scientist ranking to identify those scientists whose performance cannot be surpassed by others' with respect to all attributes. This technique can be used as a tool for short-listing distinguished researchers in case of award nomination.

1. Introduction

The field of scientometric ranking has a long history starting with the introduction of the Garfield's famous Impact Factor (Garfield, 1955) for journal ranking, and continuing with recent indices that quantify an individual's performance such as the very popular *h*-index (Bornmann, Mutz, Hug, & Daniel, 2014; Hirsch, 2005). We focus here on the family of indices that use a single number to measure a scientist's performance. Members of this family are some straightforward measures such as the average and total number of citations, the number of citations in the elite set of articles (Vinkler, 2011), variations of the *h*-index, such as the contemporary index (Sidiropoulos, Katsaros, & Manolopoulos, 2007), the *e*-index (Zhang, 2009), the *f* index (Katsaros, Akritidis, & Bozanis, 2009) and many more (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Wildgaard, Schneider, & Larsen, 2014). Their advantages and disadvantages have been documented in various studies, and the overall conclusion is that each one focuses on one (or more) but not all of the aspects of an individual's performance (project ACUMEN Wildgaard et al., 2014). For instance, the *h*-index is a proxy for the cumulative impact and productivity achievement, the contemporary *h*-index takes into account how contemporary the articles that comprise the *h*-index, the *e*-index complements the *h*-index by accounting for the ignored excess citations, etc. Therefore, it becomes evident that a fair evaluation of a scientist's work based on quantitative data must take into account multiple, uncorrelated indicators (Bornmann, Mutz, Hug, & Daniel, 2011).

* Corresponding author at: Electrical and Computer Engineering Department, University of Thessaly, Greece. Tel.: +30 2421074975.
E-mail addresses: asidirop@gmail.com (A. Sidiropoulos), agogoglou@csd.auth.gr (A. Gogoglou), dkatsar@inf.uth.gr (D. Katsaros), manolopod@csd.auth.gr (Y. Manolopoulos).

[1] Work done while the author was on sabbatical leave at Yale University.

http://dx.doi.org/10.1016/j.joi.2016.04.009
1751-1577/

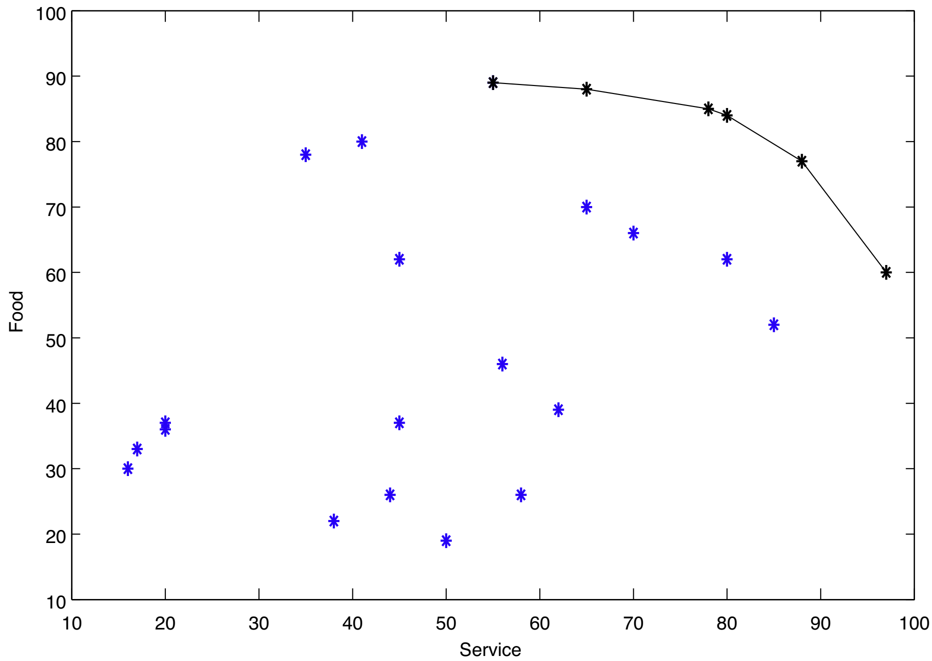

Fig. 1. Skyline plot for restaurants.

A straightforward way to do this is to define a set of weights for the set of indices and compute a weighted average score. A precondition for this process is the normalization of the scores in such a way that they are comparable. This is usually impossible, since most of the indices are not upper-limited. In addition to that, the definition of the corresponding weights will certainly be arbitrary. An alternative is to address the problem as a "rank aggregation" problem (Fagin, Lotem, & Naor, 2001; Langville & Meyer, 2014), and fuse the ranking lists produced by each indicator by appropriately adapting methods such as those reported in Tsai (2014). Still, the selection of the fusion algorithm will raise questions about its appropriateness and fairness.

We argue here that we do not need to produce a single ranked list; we should simply identify those individuals that have not been surpassed by than any other individual with respect to all considered indicators. This is the concept described as *skyline*, and calculated by the respective operator (Börzsönyi, Kossmann, & Stocker, 2001). The resulting set of distinguished individuals is called the *skyline set*.

We will explain how this works by presenting an example from Chomicki, Godfrey, Gryz, and Liang (2003). Assuming that we want to choose a restaurant based on its service and food quality. Having a rating for each restaurant's service and their food quality we can produce two rank tables, one for each evaluation metric (service,food). It is difficult to produce a global rank table by combining the existing two. That is because we cannot define the relation between service and food. Attempts to define such a relation will be prove to be arbitrary. The skyline set notion enables us to detect the best restaurants (given the attributes) by combining the two metrics (or more metrics). The skyline set consists of the set of restaurants (generally the set of objects) none of which can be surpassed with regards to any of the attributes by any other restaurant (or object in general). A geometrical view is shown in Fig. 1. In this plot every point represents a restaurant (an object). The coordinates of each object are defined by the score of the object for each metric. Each metric corresponds to one dimension. Since the higher the score of the two metrics (service and food) the better the object (restaurant), the objects that can surpass all the other objects are distinguished as the top choices. A two metrics rank can be presented with a 2D plot.

2. Definition and calculation of a dataset's skyline set

In this section we will present the original definition of skyline set and a basic, efficient algorithm for its computation, as presented by Borzsonyi adding the mathematical notation.

Definition 1 *(Dominance relationship).* Given two multidimensional points s_1 and s_2 with attributes (dimensions) α from a space D, if s_1 is equal to or better than s_2 in all dimensions, and s_1 is better than s_2 in at least one attribute, we would say that s_1 dominates s_2 and write $s_1 \succ s_2$. That is:

$$s_1 \succ s_2 : (\forall \alpha \in D, s_1.\alpha \geq s_2.\alpha) \wedge (\exists \alpha \in D, s_1.\alpha > s_2.\alpha).$$

Definition 2 *(Skyline set).* The skyline set comprises the set of points not dominated by any other point.