Available online at www.sciencedirect.com

ScienceDirect

Journal of Informetrics 1 (2007) 26–34

Journal of INFORMETRICS
An International Journal

www.elsevier.com/locate/joi

Some measures for comparing citation databases

Judit Bar-Ilan[a,*], Mark Levene[b], Ayelet Lin[a]

[a] *Department of Information Science, Bar-Ilan University, Ramat Gan 52900, Israel*
[b] *School of Computer Science and Information Systems, Birkbeck University of London, Malet Street, London WC1E 7HX, UK*

Received 29 June 2006; received in revised form 3 August 2006; accepted 3 August 2006

Abstract

Citation analysis was traditionally based on data from the ISI Citation indexes. Now with the appearance of Scopus, and with the free citation tool Google Scholar methods and measures are need for comparing these tools. In this paper we propose a set of measures for computing the similarity between rankings induced by ordering the retrieved publications in decreasing order of the number of citations as reported by the specific tools. The applicability of these measures is demonstrated and the results show high similarities between the rankings of the ISI Web of Science and Scopus and lower similarities between Google Scholar and the other tools.

Keywords: Similarity measures; Rankings; Citation databases

1. Introduction

Citation analysis is a major subfield of informetrics. Until recently the only comprehensive tool for carrying out empirical research in this area was the ISI Citation Indexes (see for example White's (2001) discussion on CAMEOs). This situation has changed, at first in individual disciplines (like CiteSeer in computer science), and now with the introduction of Elsevier's Scopus and Google Scholar.

Citation data is heavily influenced by the coverage of the specific database, since it can take into account only citations from items indexed by it. The three major tools: Web of Science (the Web version of the ISI Citation Indexes), Scopus and Google Scholar were compared and reviewed in several publications from different aspects (for example: Bauer & Bakkalbasi, 2005; Deis & Goodman, 2005; Jacso, 2005a, 2005b; Noruzi, 2005; Bar-Ilan, 2006). CiteSeer and SCISearch (a different interface of the ISI Science Citation Index) were compared by Goodrum, McCain, Lawrence, and Giles (2001). The above-mentioned studies provided numbers and descriptive statistics as a means for comparing between the different tools.

With the existence of multiple citation databases it becomes necessary to compare them systematically both from the scientometric and the informetric points of view. Descriptive statistics and specific examples are not sufficient for systematic comparison of the different citation databases. In this paper we introduce a set of measures for comparing the different citation databases. The measures compute the similarities between the rankings induced by the number of citations a publication receives in the specific database (i.e. the most cited item is ranked number 1, the second most

[*] Corresponding author. Tel.: +972 523667326; fax: +972 3 5353937.
E-mail addresses: barilaj@mail.biu.ac.il (J. Bar-Ilan), M.Levene@dcs.bbk.ac.uk (M. Levene), lineyal@netvision.net.il (A. Lin).

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.08.001

cited is ranked number 2, etc.). The use of these measures and statistical analysis of the results is demonstrated on a subset of the highly cited Israeli researchers, as defined in ISI's Highly Cited database (ISI HighlyCited.com, 2002) supplemented by the three recent Israeli Nobel prize winners.

The measures are defined in Section 2, the data collection and empirical settings appear in Section 3. In Section 4 the results are displayed and analyzed, and Section 5 concludes the paper.

2. The measures

The rankings were compared using four basic measures that complement each other. In this section the measures are defined. Each of the measures is defined for a pair of databases (A and B), where A and B can WoS (Web of Science), Scopus or Google Scholar. The measures introduced here were applied to comparing rankings of search engine rankings (Bar-Ilan, Mat-Hassan, & Levene, 2006; Bar-Ilan, Levene, & Mat-Hassan, 2006; Bar-Ilan, Keenoy, Yaari, & Levene, submitted for publication).

2.1. *Overlap and footrule*

Overlap (O) is defined as follows:

$$O = \frac{|\mathrm{PUBL_A} \cap \mathrm{PUBL_B}|}{|\mathrm{PUBL_A} \cup \mathrm{PUBL_B}|}$$

where PUBL_X is the set of publications retrieved from database X. The measure O does not take into account the rankings, it only measures the proportion of the publications retrieved from both databases out of the total number of publications retrieved by either of them.

Footrule, F, is the normalized Spearman footrule. Spearman's footrule (Diaconis & Graham, 1977; Dwork, Kumar, Naor, & Sivakumar, 2001) can be computed for two permutations, and thus it can be applied only for the publications that are ranked in both databases. Each such publication is given its relative rank in the set of publications retrieved from both databases. Suppose for the moment that there are no ties in the rankings (i.e. no two publications receive exactly the same number of items). This is an unrealistic assumption and we will deal with it in Section 3. The result of the re-rankings is two permutations σ_1 and σ_2 on $1 \ldots Z$ where $|Z|$ is the number of overlapping publications. After these transformations Spearman's footrule is computed as

$$Fr^{|Z|}(\sigma_1, \sigma_2) = \sum_{i=1}^{|Z|} |(\sigma_1(i) - \sigma_2(i))|$$

When the two rankings are identical on the set Z, $Fr^{|Z|}$ is zero, and its maximum value is $|Z|^2$ when $|Z|$ is even, and $(|Z|+1)(|Z|-1)$ when $|Z|$ is odd. When the result is divided by its maximum value, $Fr^{|Z|}$ will be between 0 and 1, independent of the size of the overlap. This measure is undefined for $|Z| = 0{,}1$. Thus we compute the *normalized Spearman's footrule*, *NFr*, for $|Z| > 1$

$$NFr = \frac{Fr^{(|Z|)}}{\max Fr^{(|Z)}}$$

NFr ranges between 0 and 1; it attains the value 0 when the relative ranking of the publications in the set Z is identical. Since we are interested in similarity measures, we define F as

$$F = 1 - NFr$$

The weakness of this measure is that it totally ignores the non-overlapping elements and only takes into account the relative rankings, thus for example if $|Z| = 2$, and these two publications are ranked at ranks 1 and 2 in database A, while in database B they are ranked at 9 and 10 (and the first eight publications are not ranked in database A), the value of F will be 1, just like the case where both A and B rank these two publications at ranks 1 and 2, respectively.