Journal of Informetrics 12 (2018) 631–655

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

How to evaluate rankings of academic entities using test data

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[b]

[a] Media Lab, Stellenbosch University, 7602 Matieland, South Africa
[b] Department of Computer Science, Stellenbosch University, 7602 Matieland, South Africa

ARTICLE INFO

Article history:
Received 17 April 2018
Received in revised form 1 June 2018
Accepted 7 June 2018

Keywords:
Evaluating rankings
Test data
Cranfield paradigm
Significance testing

ABSTRACT

In the field of scientometrics, impact indicators and ranking algorithms are frequently evaluated using unlabelled test data comprising relevant entities (e.g., papers, authors, or institutions) that are considered important. The rationale is that the higher some algorithm ranks these entities, the better its performance. To compute a performance score for an algorithm, an evaluation measure is required to translate the rank distribution of the relevant entities into a single-value performance score. Until recently, it was simply assumed that taking the average rank (of the relevant entities) is an appropriate evaluation measure when comparing ranking algorithms or fine-tuning algorithm parameters.

With this paper we propose a framework for evaluating the evaluation measures themselves. Using this framework the following questions can now be answered: (1) which evaluation measure should be chosen for an experiment, and (2) given an evaluation measure and corresponding performance scores for the algorithms under investigation, how significant are the observed performance differences?

Using two publication databases and four test data sets we demonstrate the functionality of the framework and analyse the stability and discriminative power of the most common information retrieval evaluation measures. We find that there is no clear winner and that the performance of the evaluation measures is highly dependent on the underlying data. Our results show that the average rank is indeed an adequate and stable measure. However, we also show that relatively large performance differences are required to confidently determine if one ranking algorithm is significantly superior to another. Lastly, we list alternative measures that also yield stable results and highlight measures that should not be used in this context.

1. Introduction

New metrics and indicators for scoring academic entities are frequently proposed. To evaluate indicators on their utility for some task different approaches are taken. A metrics's mathematical soundness can be validated using axiomatic approaches (Altman & Tennenholtz, 2010; Bouyssou & Marchant, 2016). Two or more indicators can be compared to each other using correlation analyses. While this can yield some insight into proposed indicators, correlation analyses are problematic on their own (Thelwall, 2016) and can only be used as a comparison to some baseline (such as citation counts used as proxy for quality).

* Corresponding author at: Media Lab, Stellenbosch University, 7602 Matieland, South Africa.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

https://doi.org/10.1016/j.joi.2018.06.002
1751-1577/© 2018 Elsevier Ltd. All rights reserved.

Another approach is to use test data to evaluate ranking algorithms. One drawback of using test data is that its collection is expensive and time consuming. To decrease this effort lists of readily available data are often used as proxies for human judgements. Examples of such lists are: researchers that have received fellowship status at learned societies in recognition of their work (Dunaiski, Geldenhuys, & Visser, 2018; Nykl, Campr, & Ježek, 2015; Nykl, Ježek, Fiala, & Dostal, 2014); researchers that have won life-time contribution or innovation awards (Dunaiski, Visser, & Geldenhuys, 2016; Fiala, 2012; Fiala, Rousselot, & Ježek, 2008; Fiala & Tutoky, 2017; Gao, Wang, Li, Zhang, & Zeng, 2016; Nykl et al., 2014); and researchers that are frequently board members of prestigious journals (Fiala, Šubelj, Žitnik, & Bajec, 2015). For paper-level rankings, best paper awards or high-impact paper awards have been used (Dunaiski et al., 2016; Dunaiski & Visser, 2012; Mariani, Medo, & Zhang, 2016; Sidiropoulos & Manolopoulos, 2005).

We use the terms *metric* and *ranking algorithm* synonymously since they assign scores to academic entities that can be converted into a ranking (sorted list of entities with ascending ranks). When using test data (a subset of all entities considered important) to evaluate a ranking, some *evaluation measure* is needed to translate the rank distribution of the relevant entities into a single-value performance score. This paper deals with the evaluation measures and how they should be applied when evaluating ranking algorithms using test data.

Frequently, conclusions are based on simply using the average rank of the relevant entities as a performance score. This evaluation measure has been used to compare ranking algorithms to each other but also to draw conclusions about properties of the internal workings of the algorithms. For example, it has been used to judge whether self-citations should be included when computing impact scores of authors (Dunaiski et al., 2018; Nykl et al., 2014). Using the average rank as evaluation measure makes the assumption that if algorithm A ranks the important entities on average higher than algorithm B, then A must be better than B. However, it remains unknown whether the observed performance difference was obtained by algorithm A's superior ranking capabilities or was caused by outliers on a skewed rank distributions or simply occurred by chance. Moreover, how significant are the performance differences between the algorithms under investigation? Recently, alternative evaluation measures are adopted (Fiala & Tutoky, 2017) but the same problems remain: how confident are we about the obtained results?

In this paper we answer the above questions by addressing the following problem. The number of entities in a test data set is orders of magnitudes smaller than the number of authors or papers in real-world publication databases. Therefore the rank distribution of the test data entities is sparse and does not necessarily contain many high ranks. This situation causes many standard evaluation measures to become less effective. We show this by using methodologies from query-based information retrieval frameworks and adapting them for rankings of academic entities (Buckley & Voorhees, 2000; Sakai, 2006; Voorhees & Buckley, 2002).

Using our proposed framework, we analyse the discriminative power and stability of the evaluation measures on sparse rankings. The discriminative power is defined in terms of how well an evaluation measure distinguishes between significant and insignificant differences in rankings. The stability of a measure is based on its consistency of producing the correct results under changing conditions. In other words, we analyse an evaluation measure's general performance when the underlying data is changed and its volatility to rank biases.

The diagram in Fig. 1 depicts the workflow followed in this paper. Given a database of academic entities (papers or authors), they are ranked by metrics M_1 through M_k that assign scores to the entities. In Section 2 we discuss how these scores are converted into fair ranks. The next step is to extract the ranks of relevant entities of a test data set, in this case 'Test Set 1'. We describe the different test data sets used in this paper in Section 3 and outline the motivation behind this paper. Section 4 describes the most common evaluation measures in the context of academic entities and how they can be adjusted for percentile rankings. Based on the rank distribution of the relevant entities, the evaluation measures are used to compute performance scores for the metrics. We then formulate the framework of how these evaluation measures are evaluated (Section 5). In Section 6 we discuss the results of this second-order evaluation.

We make the following contributions:

- We propose a framework for evaluating evaluation measures based on rankings of academic entities that are part of test data. Using this framework the stability and discriminative power of the most common evaluation measures are analysed.
- The proposed methodology provides the capability of computing significance levels associated with performance differences between ranking algorithms.
- We show that simple measures such as the average or median rank have high discriminative power and are stable evaluation measures.
- We find that using permille rankings does not improve the performance of evaluation measures in general except for the nDCG measure which should only be used with permille rankings.
- Our results show that a "one size fits all" evaluation measure does not exist and that appropriate measures have to be chosen carefully based on the underlying data.

2. Converting scores to ranks

Ranking algorithms and impact indicators usually produce scores that are associated with entities. However, scores from different metrics are not directly comparable and have to be converted to ranks first. An entity with a larger score usually indicates that it is "better" than entities with smaller scores produced by the same metric. Therefore, the output of metrics