Aggregating Productivity Indices for Ranking Researchers across Multiple Areas

Harlley Lima, Thiago H. P. Silva, Mirella M. Moro,
Rodrygo L. T. Santos, Wagner Meira Jr., Alberto H. F. Laender
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{harlley,thps,mirella,rodrygo,meira,laender}@dcc.ufmg.br

ABSTRACT

The impact of scientific research has traditionally been quantified using productivity indices such as the well-known h-index. On the other hand, different research fields—in fact, even different research areas within a single field—may have very different publishing patterns, which may not be well described by a single, global index. In this paper, we argue that productivity indices should account for the singularities of the publication patterns of different research areas, in order to produce an unbiased assessment of the impact of scientific research. Inspired by ranking aggregation approaches in distributed information retrieval, we propose a novel approach for ranking researchers across multiple research areas. Our approach is generic and produces cross-area versions of any global productivity index, such as the volume of publications, citation count and even the h-index. Our thorough evaluation considering multiple areas within the broad field of Computer Science shows that our cross-area indices outperform their global counterparts when assessed against the official ranking produced by CNPq, the Brazilian National Research Council for Scientific and Technological Development. As a result, this paper contributes a valuable mechanism to support the decisions of funding bodies and research agencies, for example, in any research assessment effort.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Research performance; Bibliometric indicators; Ranking aggregation; Cross-disciplinarity

1. INTRODUCTION

Evaluating a group of researchers is a permanent problem within research and academic institutions, laboratories and funding agencies. Usually, this process involves forming highly qualified committees that must meet, define evaluation criteria and perform the evaluation. Moreover, it is very costly in terms of time, because evaluating numerous researchers (their curricula and publications) is not a simple process.

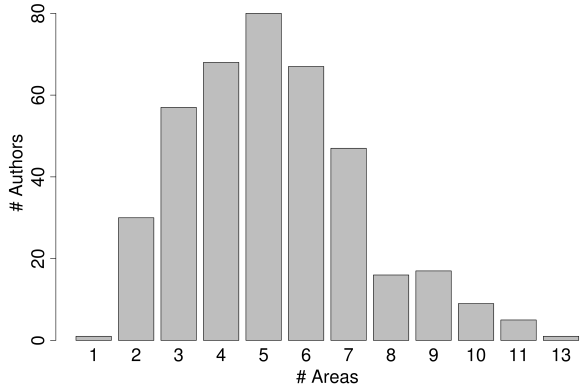


Figure 1: Distribution of Brazilian CS researchers per number of active areas. A researcher is deemed active in an area if at least 10% of the researcher's publications have been classified in the area.

The decision of which researchers should be at the top (for hiring, promoting, funding, or distributing grants, scholarships, awards and so on) is typically based on criteria such as number of publications, impact of publications, number of undergraduate and graduate students under supervision, number of advised MSc and PhD theses, and participation in committees (conferences, journal editorial boards, technical committees, etc). Clearly, the effectiveness of the resulting ranking depends on how each criterion is assessed and the period of time covered in the assessment [1, 23]. Although a researcher's performance cannot be measured solely by bibliometric indices, such indices have become widely used to measure the productivity of researchers. Examples include the number of citations, h-index [16], g-index [11] and citation z-score [22]. Likewise, most academic search platforms (such as ArnetMiner,[1] Google Scholar[2] and Microsoft Academic Search[3]) use some of such indices to rank researchers. However, a common limitation of global bibliometric indices

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
Copyright 2013 ACM 978-1-4503-2077-1/13/07 ...$15.00.

[1] http://arnetminer.org/
[2] http://scholar.google.com/
[3] http://academic.research.microsoft.com/

is that they do not account for the dynamics of different areas when assessing a researcher's performance.

Ranking researchers without regarding the specificity of different areas is arguably unfair and potentially error-prone. For instance, consider the area of Human-Computer Interaction within the broad field of Computer Science (CS) [2]. Experimental evaluation in this area usually takes more time than in other CS areas when arranging and assessing users' feedback is necessary. On the other hand, CS areas such as Databases and Computer Graphics do not usually face the same problem because their experimental evaluations depend on assessing the outcome of an automatic process, such as a query evaluation or a graphics rendering engine. Likewise, researchers from some areas may have fewer publications than others, but with a potentially higher impact in their community. To aggravate the problem, many researchers publish in more than one area. To illustrate this observation, Figure 1 shows the distribution of Brazilian CS researchers per number of areas where they have at least 10% of their publications.[4] The figure shows a normal distribution with an average of 5 active areas per researcher. Most notably, almost all researchers are active in more than one area, with a few being active in up to 13 areas.

In order to improve the assessment of academic research, we introduce a novel approach for ranking researchers across multiple—and potentially distinct—research areas within an academic field. In particular, our approach estimates the performance of a researcher in each area relatively to the performance of other researchers in the same area, so as to ensure that the specificity of each area is accounted for accordingly. The researcher's relative performance in multiple areas is then aggregated into a unified ranking, by projecting the performance in each individual area to the corresponding performance in the researcher's base area, i.e., the area in which the researcher has most of her academic production. Our approach is generic and can be used to leverage different criteria for ranking researchers across areas. In particular, to demonstrate the feasibility of our approach, we devise cross-area rankings based on three different criteria: publication volume, number of citations and h-index. Our thorough evaluation considering multiple areas of researchers shows that our cross-area indices outperform their global counterparts when assessed against an official ranking produced by CNPq[5], the Brazilian National Research Council for Scientific and Technological Development.

The contributions of this paper are three-fold:

1. We demonstrate the limitation of a one-size-fits-all approach for ranking researchers across multiple areas, using the Brazilian CS community as a case study;

2. We introduce a novel ranking aggregation approach to rank researchers across multiple areas, which respects the idiosyncrasies of different areas;

3. We thoroughly evaluate the proposed approach applied to three well-known productivity indices: publication volume, citation count and h-index.

In the remainder of this paper, Section 2 provides background on related approaches to assess academic research. Section 3 introduces our novel approach for ranking researchers across multiple areas. In turn, Sections 4 and 5 describe the experimental methodology and the results of the evaluation of our approach, respectively. Lastly, Section 6 discusses our conclusions and directions for future research.

[4]More details about the dataset used to generate the statistics in Figure 1 will be given later in Section 4.
[5]http://www.cnpq.br

2. RELATED WORK

A classic problem in science consists in ranking scientists from distinct academic fields, such as mathematics, physics, and CS [5, 23, 24]. For instance, Podlubny [24] proposed an equivalence table for citations in different academic fields based upon the continued observation of the citation rates in these fields. After normalization, this table established, for instance, that one citation in mathematics corresponded roughly to 5 citations in engineering, 19 in physics, and 78 in the biomedical field, among others. However, the ad hoc nature of this study requires a periodical correction of the equivalence table over time, which is further aggravated by the unstable evolution of individual fields. Likewise, Radicchi et al. [25] performed an empirical analysis of the distribution of citations for publication among research fields. They proposed a relative indicator $c_f = c/c_o$, where c is the publication citation number and c_o is the average number of citations per article for the scientific field, which rescaled the distributions of citation for publication in different scientific fields upon the same curve when c_f is applied. Later, Bornmann and Daniel [4] explored the advantages of such indicator for the area of chemistry.

Claro and Costa [9] proposed the x-index as a cross-field bibliometric indicator in order to compare researchers from distinct scientific fields. The proposed index considers the top authors in each field (based upon the publication volume) as a reference set for the field. The productivity of the remaining authors in each field is then computed relatively to the field's reference set, so as to enable a cross-field comparison. A limitation of this approach is that the reference authors may be actually outliers, since publication volume is not necessarily an indicator of publication quality.

Another bibliometric indicator that tries to reduce the possible discrepancies between scientific fields is the crown index [28]. In particular, this index builds upon the observation that the average number of citations per publication varies across fields [21, 24]. To exploit this observation, the crown index normalizes a researcher's citation count by the expected number of citations in each field. Lundberg [22] extended this idea by normalizing the citation count at the publication (as opposed to the researcher) level. An additional extension used a logarithm-based normalization and assigned a weight to each publication according to a skewed distribution of citations over publications.

Freire and Figueiredo [14] proposed a productivity index to rank individuals within a target group (as opposed to ranking all individuals) in a collaboration network using solely the relationships among these individuals. In particular, the importance assigned to a specific individual by this index was proportional to the intensity of his or her collaboration with other individuals outside the target group. In the same work, the authors also considered a variation of their index to rank entire groups of individuals and applied it to rank graduate programs in the CS field.

In online rankings, such as the ones provided by Microsoft Academic Search and ArnetMiner, the top authors are classified according to their h-index by default. The h-index of a

researcher is defined as the largest number h for which the researcher has h publications with at least h citations each [16]. The main disadvantage of solely applying the h-index is that it favors researchers with long academic careers [6]. This index is also insensitive to authors with many papers with few or no citations, and those with few papers but many citations. In order to overcome these limitations, the h-index has also been adapted and combined with other indices to produce improved researcher rankings [11, 13, 16, 23]. In this vein, Bollen et al. [3] conducted an evaluation of different rankings by using 39 impact indicators and concluded that the concept of scientific impact is multi-dimensional and cannot be measured by using only one indicator.

Regarding cross-area ranking, according to Glänzel and Schubert [15] and Oliveira et al. [23], the performance of a researcher may significantly vary depending on his or her areas of activity. Furthermore, the impact factor of a journal[6] in basic and fundamental areas usually receive higher values than in specialized or applied areas [2]. In other words, bibliometric indicators provide a relative comparison that needs to be adapted or combined with other indicators as well as with additional information to cover the singularities of the desired classification. Despite some recent efforts to provide indices to enable the comparison of researchers across distinct fields [9, 15, 28], to best of our knowledge, our proposed approach is the first attempt to produce an aggregated ranking of researchers based upon their performance across distinct areas within the same field.

3. RANKING RESEARCHERS ACROSS MULTIPLE RESEARCH AREAS

Assessing the impact of scientific research is a challenging task, which has led to several productivity indices in the past, as discussed in the previous section. Nonetheless, the majority of these indices disregards the specificity of different research areas and assesses researchers based on global statistics, such as each researcher's publication volume and citation count. As these statistics may vary substantially across different areas, such indices may be unsuitable in a cross-area research evaluation effort. To illustrate this observation, Table 1 summarizes the distribution of publication volume and citations for researchers in the 23 CS areas represented in the dataset used in our evaluation, as described later in Section 4.1.

As observed from Table 1, different research areas have different publishing targets, with some focusing on conferences more than journals, and others doing the other way around. More importantly, researchers in different areas show clearly distinct publishing patterns, with the average volume and number of citations per researcher varying greatly within each area and across multiple areas. In order to provide an unbiased assessment of a researcher's productivity across multiple research areas, we argue that a productivity index should possess the following properties:

- *Plurality.* The productivity of a researcher should be assessed in all areas in which the researcher has published.
- *Diversity.* The profile of each research area should be considered when assessing a researcher's productivity.
- *Equality.* All research areas should be regarded as equally important and deserving of scientific merit.

To overcome the limitation of existing productivity indices and address the above requirements, we introduce a novel approach for ranking researchers across multiple research areas, by aggregating evidence of each researcher's productivity in each individual area. Inspired by ranking aggregation approaches in distributed information retrieval [7], our approach is generic and can be used to leverage any existing productivity index (e.g., publication volume and citation count, and even the well-known h-index) as evidence for ranking researchers across areas. In particular, our approach aims to produce a global ranking of researchers in a given field by tracking the position of each researcher in the rankings produced for the various research areas within that field. By doing so, we recognize the *plurality* of a researcher's publishing pattern, as previously demonstrated in Figure 1.

To formalize our approach, let s_i^a be the score assigned to researcher i with respect to a given research area a. This score is calculated adding up the contribution of each of the researcher's publications to the area a. To this end, we assume that each publication can be classified into multiple areas, based upon, e.g., the areas of interest of the venue where the publication appeared. Under this assumption, score s_i^a is defined as:

$$s_i^a = \sum_{j=1}^{n_i} \mathbf{1}_j^a \frac{s_{i,j}}{m_{i,j}}, \quad (1)$$

where n_i is researcher i's total number of publications, $\mathbf{1}_j^a$ is an indicator function (which is 1 if the researcher's j-th publication covers the area a and 0 otherwise), $s_{i,j}$ is the score conferred by this publication to researcher i,[7] and $m_{i,j}$ is the total number of areas covered by this publication. The latter quantity acts as a normalization factor, so as to ensure that the score $s_{i,j}$ is not accounted for more than once when aggregating the researcher's scores across multiple areas.

Given the aforementioned discrepancies between areas, as illustrated in Table 1, the area scores produced for a researcher may not be comparable across areas. In order to ensure that the *diversity* of areas is respected, we propose to use the percentile rank p_i^a of researcher i in each area a instead of the researcher's raw area score s_i^a. As a result, we consider the position of the researcher relatively to other researchers in the area. To formalize this intuition, we estimate the percentile rank p_i^a as:

$$p_i^a = \frac{l_i^a + 0.5e_i^a}{N^a}, \quad (2)$$

where N^a is the total number of active researchers in area a, l_i^a and e_i^a are the number of such researchers with a score lower than or equal to that of researcher i, respectively. The latter quantity ensures that tied researcher scores are accounted for appropriately [17].

Lastly, we must ensure that the *equality* between areas is respected. To this end, we introduce a novel ranking aggregation approach, which enforces an equal treatment between areas. In particular, our approach equates the percentile rank attained by a researcher in a particular area to

[6] `http://thomsonreuters.com/products_services/science/free/essays/impact_factor/`

[7] For a volume-based score, we have $s_{i,j} = 1, \forall j \in [1, n_i]$; for a citation-based score, $s_{i,j}$ is the total number of citations of researcher i's j-th publication.

the same percentile rank in other areas. Formally, let b_i denote the base area of researcher i, i.e., the area in which the researcher has the highest percentile rank p_i^a, according to:

$$b_i = \arg\max_{a \in A} p_i^a, \quad (3)$$

where A represents the set of all areas under consideration within a field. In order to aggregate the scores attained by a researcher across multiple areas, we first project the percentile rank attained by the researcher in each of her areas of activity onto her base area. As a result, we can further reward the researcher with the base area score corresponding to her percentile ranks in multiple areas. For example, a researcher with a percentile rank 90 in Databases and 30 in Computer Graphics should be rewarded with the scores corresponding to the 90 and 30 percentiles in Databases, assuming that this is the researcher's base area. Formally, the aggregated, cross-area score s_i of researcher i in light of all areas $a \in A$ in a field of interest is given by:

$$s_i = \sum_{a \in A} \frac{f_b(p_{i,a})}{f_b(1.0)}, \quad (4)$$

where the projection function $f_b(p_{i,a})$ maps the percentile rank $p_{i,a}$ attained by researcher i in area a to the score corresponding to this percentile in the researcher's base area b, and $f_b(1.0)$ returns the maximum score in b among all researchers in this area. This normalization step eliminates any bias towards researchers with a prolific base area (e.g., an area with a profile of high publication volume and citations), further enforcing the equality between areas.

4. EXPERIMENTAL METHODOLOGY

This section discusses the methodology underlying the experiments described in Section 5 for the evaluation of our proposed approach to rank researchers across areas. In particular, we aim to answer the following research questions:

Q1. Can we improve existing productivity indices with our cross-area ranking approach?

Q2. Can we effectively combine global and cross-area indices for an improved ranking?

Q3. Which other factors may impact our cross-area ranking approach?

In the remainder of this section, Section 4.1 describes the dataset used to support our investigation.The baseline indices derived from this dataset are discussed in Section 4.2. Lastly, Section 4.3 discusses the procedure for evaluating these baselines as well as our proposed ranking approach.

4.1 Publication Dataset

High-quality bibliometric indices typically depend on a high-quality dataset of publications. Digital libraries, such as DBLP and IEEE Xplore provide information about the authors, title, venue, and year of a publication. However, they do not provide the citation count of each publication or the h-index of each author. Then, online research-oriented search engines (such as Google Scholar, ArnetMiner and Microsoft Academic Search) do provide h-index and other bibliometric indices but their organization by author is cumbersome. Specifically, the publications of each author are not properly grouped due to name ambiguity: the same author may appear with distinct names (synonyms), or distinct authors may have similar names (polysems) [12], thus causing split and mixed citations [20]. Furthermore, the collection indexed by such search engines may comprise everything that a person has ever authored, including non-scientific pieces, technical reports and web pages.

Therefore, we need to build a dataset with the following features: it includes only qualified publications which appeared in conference proceedings and journals; it correctly groups publications and its authors; it provides citation figures (which allows to calculate the h-index for each author); it allows to identify the area (or areas) of each publication. To do so, the next sections explain: the original data source used as base for this complex dataset (Section 4.1.1), its expansion to include journal articles (Section 4.1.2), how it was disambiguated (Section 4.1.3), and how its publications were classified into areas (Section 4.1.4).

4.1.1 Data Source for Conference Publications

We got an initial dataset from the SHINE (Simple H-INdex Estimator) project[8], which collected conference publications based on a list of venues provided by the Brazilian Computer Society's (SBC) Special Interest Groups. Each SBC SIG provided the list of conferences that cover its topics of interest. The dataset then aggregates one set of conferences for each CS area. Specifically, the SHINE dataset contains more than 800,000 publications from approximately 1,800 conferences, covering 23 CS areas, and 7.5 millions of citations, as collected in the beginning of 2011. We notice that each conference may have been suggested by more than one SIG.Table 1 shows the coverage for each of the 23 CS areas, in which the average for all areas is 88%. For 19 of those areas, SHINE covers more than 80% of the conferences from the reference list (suggested by SBC SIGs). For 13 areas, the coverage is above 90%, and for only one area the coverage is smaller than 70%.

4.1.2 Adding Journal Publications

The SHINE dataset covers only publications from conferences. In order to get publications from journals, we first obtained the list of all CS journals ranked by Qualis[9]. Specifically, Qualis is an initiative of CAPES (the Brazilian Ministry of Education agency in charge of evaluating all graduate programs in Brazil, among other goals) for rating publication venues. In order to get the complete article references, we merged the list of journal titles with the DBLP XML dataset by using the International Standard Serial Number (ISSN) or, in the absence of ISSN, the journal title. With this effort, the original SHINE dataset was expanded with 271,000 articles from 188 journals.

4.1.3 Author Name Disambiguation

Having all bibliographic data of conference papers and journal articles into one dataset, the next step is to identify the researchers and group their publications accordingly. In this step, we had to deal with the ambiguity among the author names existing in the dataset. Hence, we have applied a state-of-art Heuristic-Based Hierarchical Method for name disambiguation [10]. This method works based on the

[8] http://shine.icomp.ufam.edu.br
[9] http://qualis.capes.gov.br

Table 1: Number of covered conferences and journals and summary statistics (average, standard deviation, and median) of the distributions of publication volume and citations for researchers in 23 CS areas.

			Volume			Citations		
Research Area	#Conf.	#Jour.	Avg.	SD.	Med.	Avg.	SD.	Med.
Algorithms and Theory	354	188	10.65	9.97	8	95.98	151.95	41
Artificial Intelligence	264	163	9.82	13.24	5	72.56	119.90	26
Collaboration Systems	10	14	8.25	16.02	2	49.00	101.16	6
Computational Biology	25	28	2.55	2.53	2	14.73	19.17	7
Computer Graphics and Image Processing	108	105	9.95	10.97	5	58.34	115.63	13
Computer Networks and Distributed Systems	297	161	13.46	18.77	6	84.95	177.22	29
Computer Science Education	35	37	3.25	5.04	1	10.49	23.45	1
Databases	184	127	8.55	13.99	4	73.30	182.51	12
Fault Tolerant Systems	32	7	2.45	3.02	1	23.35	54.80	5
Formalism	49	68	2.67	4.13	1	17.51	32.93	4
Game and Entertainment	17	6	2.37	3.06	1	13.97	56.92	0
Geoinformatics	14	11	4.59	7.92	2	20.95	42.42	2
Hardware, Architecture and Embedded Systems	124	112	5.61	12.36	2	34.93	122.71	5
Health Informatics	25	67	3.56	5.83	2	13.70	34.21	3
Human-Computer Interaction	21	31	2.71	3.19	1	17.48	26.15	6
Information Systems	487	188	22.02	19.15	17	160.01	230.76	75.5
Music Computing	15	6	2.59	3.28	2	4.21	8.21	0
Natural Language Processing	59	43	4.26	4.41	2	37.80	95.34	7
Neural Networks	84	82	7.83	12.10	4	40.52	77.37	14
Programming Languages	56	23	3.81	4.46	2	47.23	104.88	8
Robotics	56	63	3.24	3.80	2	29.71	70.27	4
Security	100	98	10.05	12.81	4	31.61	142.57	3
Software Engineering	95	42	7.72	12.71	3	57.98	144.89	13

similarity of the usual citation information such as work and venue titles, and considers the coauthorship network as well.

4.1.4 Classification into Areas

The final step is to classify each publication according to one (or more) CS area. Given that the conference papers were already classified (as given by the SBC SIGs), all journal articles had also to be classified into the same 23 areas. To do so, we applied the LAC (Lazy Associative Classifier) algorithm [27], which uses associative rules to classify items. As one journal article could potentially be associated to more than one area, we have employed the multi-label classification version of the LAC algorithm [26].

Specifically, we used the already classified conference papers as the training set. Then, we used the journal articles as the test set. The classification explored title and venue as main features. LAC estimates the probability of each instance being classified in each class. Therefore, we have specified a minimum threshold to select the areas of each article. This threshold was tuned so that the distribution in the test set matched that of the training set. As a result: the average number of articles in each area is 24,424; the maximum number of articles is in *Algorithms and Theory* with 172,799; and the minimum in *Music Computing* with only 7 articles. This number also follows the ratio from Table 1.

4.2 Ranking Baselines

Having the dataset, our next step is to define which indexes (or metrics) we will compare our approach against. We have chosen the same ones from academic social platforms (such as ArnetMiner, Microsoft Academic Search and Google Scholar): volume of publications, citation count and h-index. In volume of publications, the researchers are sorted in descending order of the total of their publications. In citation count, the researchers are sorted in descending order of the sum of all citations received by its publications. In h-index, the researchers are sorted in descending order of their h-Index [16].

4.3 Evaluation Procedure

In the following, Section 4.3.1 defines the ground-truth used in our evaluation, whereas Section 4.3.2 describes the evaluation metric used to report our experimental results.

4.3.1 Evaluation Ground-Truth

The first step in this experimental setup is defining the ground truth to which our result will be compared. Since there is no world wide ranking for computer scientists, we have decided to use a real ranking for Brazilians only. Specifically, each year, CNPq distributes research fellowships for researchers from all fields. To do so, CNPq bases its decisions on reports from special committees created for each field, where each committee is responsible for evaluating all researchers who have applied for such a fellowship in that field. The fellowships are awarded in two categories. The first category comprises 4 subcategories (1A, 1B, 1C, 1D) and each of those includes a research grant. The second category (2) is the entrance one and does not include a grant. Currently 14,713 researchers from 48 different fields have such scholarships. For Computer Science, there are 406 researchers distributed over the categories as illustrated in Figure 2. Note that 32% out of the 406 researchers are attributed to category 1, and 68% to category 2.

Even though CNPq considers much more than publications in its evaluation (e.g., a research project proposal and international insertion), the CS committee emphasizes that a qualified set of publications is a fundamental requirement for success[10]. Also, given that such a rank is produced by specialized committees, it seems reasonable to consider their resulting rank as a ground truth in our experimental evaluation. To do so, we evaluated all publications of researchers

[10] `http://memoria.cnpq.br/cas/ca-cc.htm`

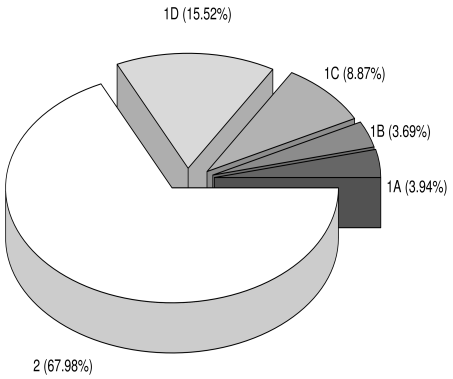


Figure 2: Distribution of CS researchers according to the ground-truth assessment provided by CNPq.

classified by CNPq in categories 1 or 2 in 2011, and disregarded all further publications from our base dataset.

4.3.2 *Evaluation Metric*

In order to evaluate our ranking approach as well as the baselines described in Section 4.2 in light of the ground-truth defined in Section 4.3.1, we use the discounted cumulative gain (DCG) metric [19]. This metric has been widely used to evaluate ranking approaches in several information retrieval tasks, most notably web search [8]. In particular, DCG adopts a non-binary notion of relevance, by assessing a given ranking based upon a graded scale, from less relevant to more relevant. In addition, this metric applies a log-based discount factor to model the fact that relevant items ranked high are preferred over the lower ranked ones [18]. Formally, the DCG at a rank position k can be defined as:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(i+1)}, \quad (5)$$

where g_i denotes the non-binary relevance grade associated with the item ranked at the i-th position. For our particular case, we define a graded relevance scale based upon each researcher's classification according to CNPq, as defined in Section 4.3.1. Specifically, we map the CNPq categories 1A, 1B, 1C, 1D and 2 to the relevance labels 5, 4, 3, 2 and 1, respectively. Finally, to bind the reported effectiveness within the interval [0,1], we use the normalized version of DCG, denoted nDCG, which is obtained by dividing the DCG@k value given by Equation (5) by the best obtainable value at the same rank cutoff k. In our experiments, we report the effectiveness of the different approaches across multiple k values, from 5 to 100, with steps of 5.

5. EXPERIMENTAL EVALUATION

In this section, we address the research questions stated in Section 4 in order to validate our proposed approach for ranking researchers across multiple scientific areas. In particular, Section 5.1 addresses research question Q1, by instantiating our ranking approach to produce cross-area versions of popular productivity indices. Section 5.2 addresses Q2, by assessing the impact of publication volume and citations as features for producing an effective ranking of researchers. Lastly, Section 5.3 addresses Q3, by performing a failure analysis of our produced rankings.

5.1 Single-Index Ranking

In order to address research question Q1, on the effectiveness of our proposed cross-area ranking approach, we contrast it to one of the most widely used productivity indices nowadays, namely, the h-index [16]. As discussed in Section 2, the h-index can be seen as a combination of the publication volume and the number of citations attained by a researcher. To obtain a cross-area version of the h-index, here called the *ca*-index ("ca" for cross-area), we propagate the global h-index from a researcher to each of his or her publications, and from each publication to the areas covered by the publication, according to Equation (1).

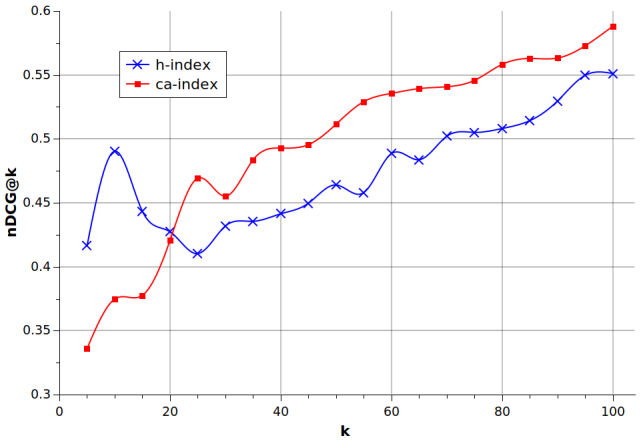
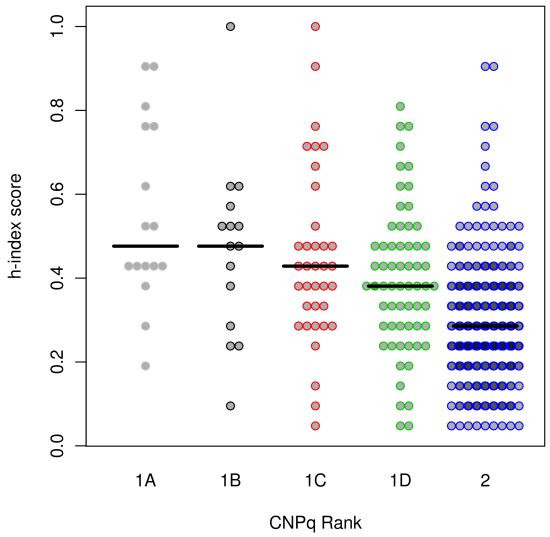


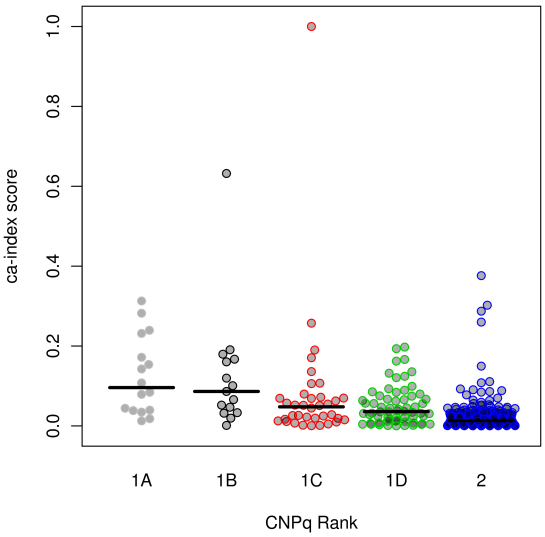
Figure 3: Comparison between researcher rankings based on the standard h-index and our ca-index.

Figure 3 compares the ranking produced by our ca-index against that produced by the original h-index, in terms of their attained nDCG@k, for a rank cutoff k varying from 5 to 100. From the figure, we first observe that our ca-index consistently outperforms the standard h-index for almost the entire range of k. The exception is for $k \leq 20$, in which case the standard h-index is the best performing. Upon further analysis, we noted that the observed difference in these top ranks is due to the presence of relevant researchers with a high specialization in only a few areas. Indeed, as shown in Figure 4, the cross-area normalization performed by ca-index reduces the dispersion of scores among the researchers classified by CNPq in each of the five considered relevance levels. Such a normalization may smooth out the performance of individual researchers, eventually missing over-specialized outliers. On the other hand, it allows for a clearer characterization of the researchers within each level, and for a better distinction between researchers across different levels, as denoted by the strictly descending median score from level 1A towards level 2 in Figure 4(b).

To further investigate the reasons behind the improvements observed in Figures 3 and 4, we analyze the impact of our cross-area ranking approach separately on the two sources of evidence underlying the h-index: publication volume and citations. Figures 5 and 6 show the results of these analyses, contrasting publication volume and citations against their counterpart cross-area indices, ca-volume and ca-citation, respectively. From Figure 5, we first observe that the estimation of publication volume is massively improved by our cross-area ranking approach. Indeed, ca-



(a) h-index score distribution



(b) ca-index score distribution

Figure 4: Distribution of (a) h-index and (b) ca-index scores across the five considered relevance levels.

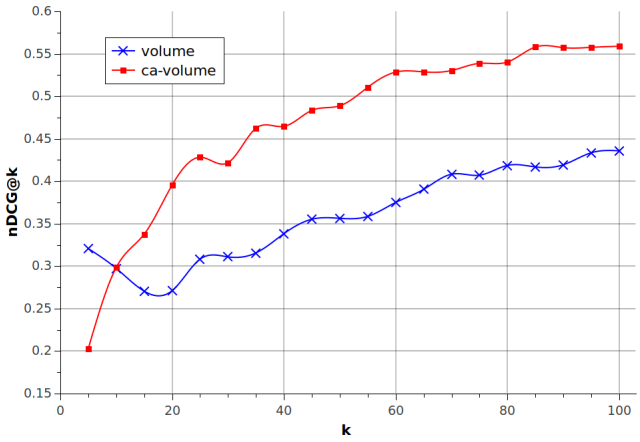


Figure 5: Comparison between researcher rankings based on publication volume and our ca-volume.

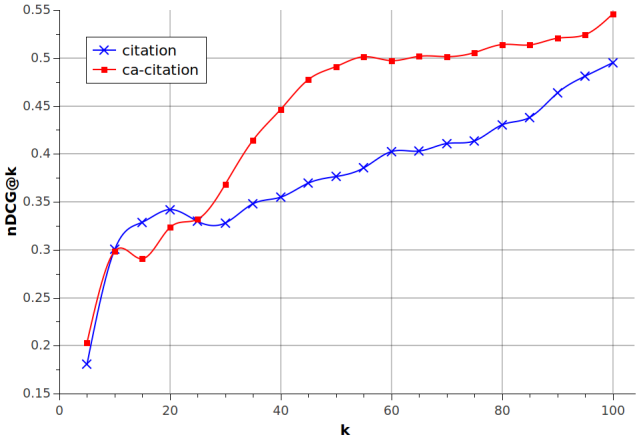


Figure 6: Comparison between researcher rankings based on citation count volume and our ca-citation.

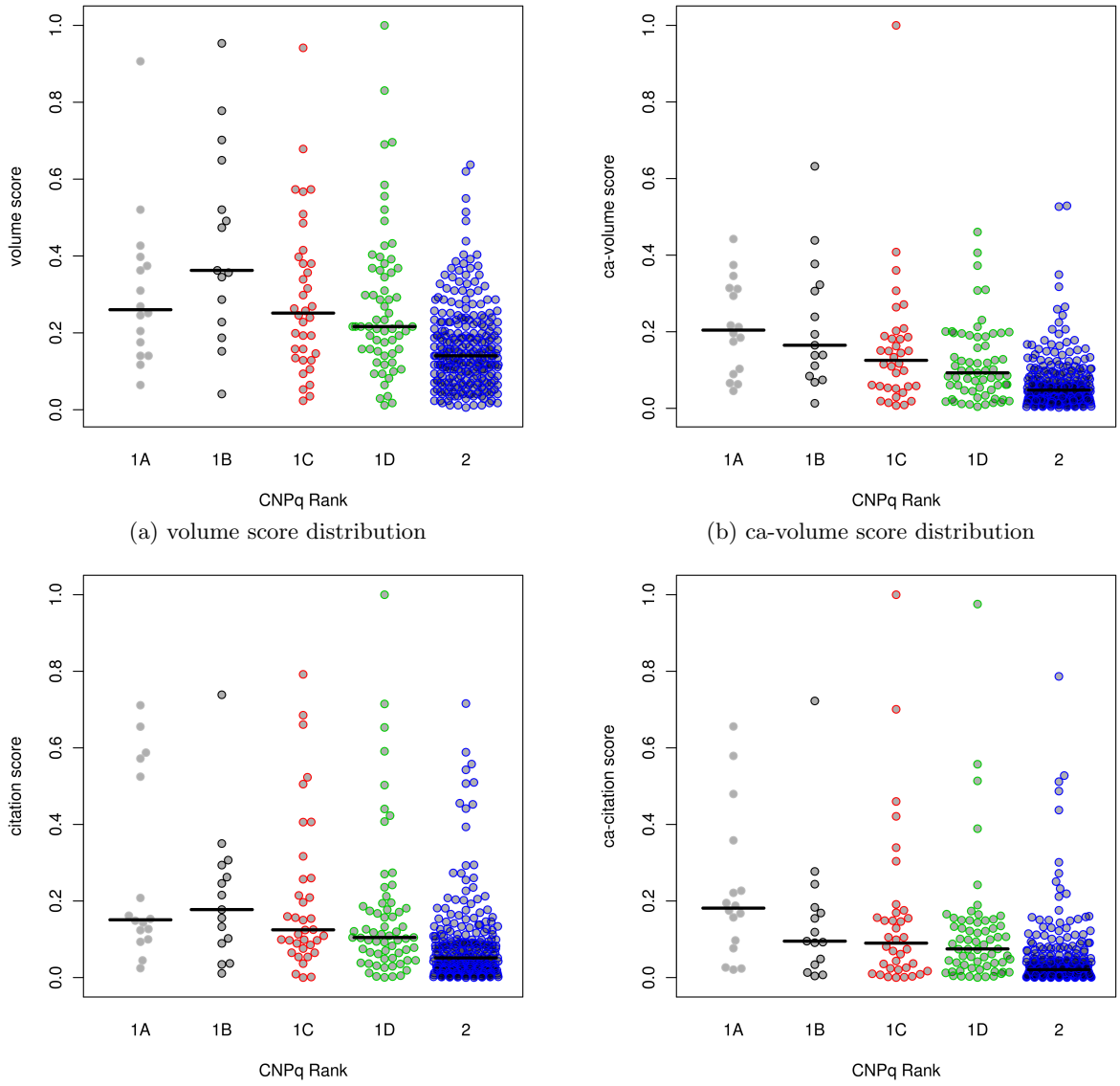
volume outperforms the standard publication volume index by a large margin from the ranking cutoff 10 onward. A similar trend is observed for ca-citation compared to the standard citation count in Figure 6, albeit with slightly less pronounced improvements compared to those observed in Figure 5.

Lastly, to further demonstrate the improvements brought by our approach for better estimating both publication volume and citation count, Figure 7 shows the dispersion of the scores produced using these two indices, prior to and after the application of our cross-area normalization. From the figure, we can observe that our approach acts as a correction for the distortion caused by simply counting volume and citations without regards to the specificity of different areas. Indeed, both ca-volume (Figure 7(b)) and ca-citation (Figure 7(d)) show a strictly descending median score from the highest rank (1A) to the lowest (2), while their global counterparts (i.e., volume in Figure 7(a) and citation in Figure 7(c)) fail to correctly align the two most distinguished categories, namely, 1A and 1B.

Overall, the results in this section answer research question Q1, by demonstrating the effectiveness of our approach for ranking researchers across multiple areas of a scientific field. Indeed, as shown in this section, our approach improves upon standard productivity indices based on publication volume and citation count, as well as upon their combination, as embodied by the well-known h-index.

5.2 Multi-Index Ranking

The results in the previous section attest the effectiveness of our cross-area ranking approach in contrast to standard productivity indices. On the other hand, the relatively higher performance attained by these global indices at early ranks suggests that they provide complementary evidence to that exploited by our approach. To investigate whether this is the case, in this section, we address research question



(a) volume score distribution

(b) ca-volume score distribution

(c) citation score distribution

(d) ca-citation score distribution

Figure 7: Distribution of (a) volume, (b) ca-volume, (c) citation, and (d) ca-citation scores across the five considered relevance levels.

Q2, by assessing the potential for combining the standard h-index with our novel ca-index. To this end, we linearly combine these two indices, giving equal weights to both in the final combination. We leave the automatic identification of optimal weights from the available data as future work.

Figure 8 shows the results of this experiment. In particular, the figure shows curves corresponding to h-index, ca-index, and their combination. From the figure, we observe that the combination h-index+ca-index performs at least as effectively as ca-index (the top performing of the two individual indices) in terms of nDCG for almost all considered ranking cutoffs. This observation answers research question Q2, by demonstrating that there is scope for further improving our ca-index, particularly at higher ranks, by combining it with other indices, such as the global h-index.

5.3 Failure Analysis

This section further analyses our ca-index ranking results in light of the ground-truth classification of Brazilian CS researchers provided by CNPq. It specifically addresses research question Q3, on which other factors may impact our index results. It also discusses how our ranking may be used for spotting researchers that may be shortlisted for promotion or demotion (of course, subject to further analysis).

In order to aid this comparison, Table 2 presents a confusion matrix, organizing the universe of researchers under consideration according to their classification by CNPq and by our ca-index approach.[11] In particular, each cell c_{ij} in

[11]The classification by ca-index is performed by splitting the score distribution produced by ca-index into five categories with the same sizes as the corresponding categories in the ground-truth produced by CNPq, as depicted in Figure 2.

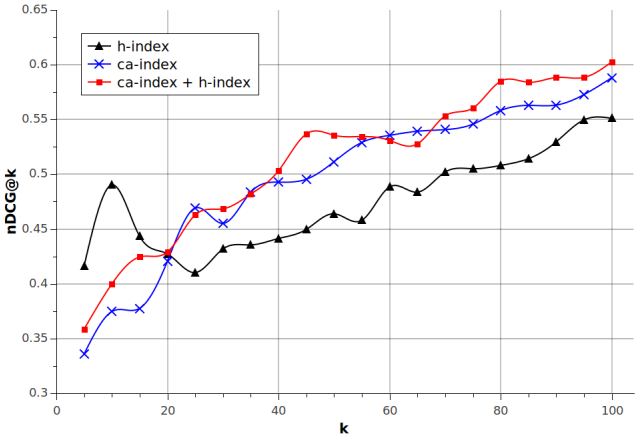


Figure 8: Comparison between researcher rankings based on the standard h-index, our ca-index, and their linear combination.

Table 2 presents the percentage of researchers classified by CNPq in the i-th category and by our ca-index in the j-th category. Note that percent figures are computed with respect to *all* considered researchers, as opposed to only those in a particular category. In other words, row 1A shows how ca-index has ranked the researchers that are currently on CNPq category 1A. For example, ca-index has classified 0.75% of CNPq researchers 1A as 1B (as shown by the first row, second column), and so on.

		ca-index				
		1A	**1B**	**1C**	**1D**	**2**
CNPq	**1A**	1.01	0.75	0.75	0.5	1.01
	1B	0.75	0.75	0.5	1.51	0.25
	1C	0.75	0.75	1.76	3.02	2.76
	1D	1.01	1.01	1.76	3.52	8.54
	2	0.5	0.5	4.27	7.04	55.03

Table 2: Confusion matrix for the rankings induced by ca-index (columns) and the CNPq classification (rows). The highlighted cells show informative discrepancies between the classification of category 1 and category 2 researchers. Percent figures are computed with respect to *all* considered researchers.

The percentage of researchers correctly classified (according to CNPq) is presented in the main diagonal (for a total or 62%). If all researchers were perfectly ranked, the main diagonal would add up to 100%, whereas the remaining cells would be 0. Overall, the values in the main diagonal are relatively high, particularly for the classification of researchers in category 2. The observed discrepancies can be explained mainly because CNPq uses other sources of evidence to rank researchers, such as their number of graduate students, contribution to innovation, research group leadership, and participation in international committees.

A similar analysis can be conducted by grouping together all researchers in the subcategories 1A-1D into a broad category 1. The major difference between CNPq categories 1 and 2 is that researchers in the former get an extra grant (besides a regular scholarship stipend). Therefore, a classification of researchers in either of these two categories can be useful for CNPq for two main reasons: to decide which researchers should receive a scholarship and which should also receive a grant. Note that the distinction between category 1 and category 2 researchers is also the most time consuming one, as most researchers fall into the latter category. In this binary classification scenario, our approach produces a correct classification for 75% of all considered researchers (20% in category 1 and 55% in category 2).

Another interesting way of interpreting the results in Table 2 is by considering that this table summarizes the distribution of scholarships related to the researchers' production. Therefore, these results may aid the decision making process by pointing out researchers that could be potentially promoted or demoted. Specifically, all researchers above the main diagonal could be further considered for a demotion, whereas those below the main diagonal could be further considered for a promotion. For example, the highlighted row at the bottom of the table represents researchers ranked by ca-index in category 1 that currently have a category 2 scholarship. Those could be further analyzed and face a "promotion" to category 1. Likewise, the highlighted column on the right represents researchers ranked in category 2 that are currently in CNPq category 1. These researchers could also be further analyzed and face a "demotion" to category 2.

Normally, such decisions should not be automatically made, because it is also important to understand why such discrepancies (for more or less) appear in Table 2. For example, the researchers at row 1A column 2 (CNPq 1A ranked by ca-index as 2) may be those who have stronger contributions other than publications. On the other hand, the researchers at row 2 column 1A (CNPq 2 ranked by ca-index as 1A) may be those who have many publications but weaker overall profiles (for example, they may not have formed PhD students yet).

6. CONCLUDING REMARKS

In this work, we have addressed the problem of ranking researchers by their scientific production. Motivated by defining a fair ranking process, our ranking is tuned to work across different research areas, a feature that distinguishes it from other previous work. To do so, our ca-index focuses on the principles of plurality, diversity, and equality of research areas. Then, it builds upon individual ranks for each area by aggregating them into one cross-area ranking.

Moreover, each year, specialized committees must evaluate hundreds of researchers for dozens of scholarships, job positions, grants, and so on. Applying ca-index quickly provides an insight on the overall production of the researchers, with the extra benefit of being tailored to work across areas. Furthermore, as illustrated by the CNPq case study, ca-index does indeed point out outliers that are worth further analysis for demotions and promotions.

Our approach was also further evaluated and compared to widely used indices such as the h-index. Our experimental evaluation has shown that ca-index outperforms the h-index when considering a comparison against our ground truth. The results for a multi-index ranking have also shown the potential for combining ca-index to other commonly used indices. As future work, we plan to expand even further our approach by considering other indicators from the researchers' profiles (including years since they have been awarded their PhD and their number of supervised students,

for example). Finally, we aim to apply the ca-index to other fields, with different characteristics.

7. ACKNOWLEDGMENTS

The authors would like to thank Thiago C. M. Salles and Itamar S. V. Hata for their invaluable help with the disambiguation and classification of the SHINE dataset, as well as the team behind the SHINE project, notably Altigran S. da Silva. The authors would also like to acknowledge their individual grants from the CNPq and FAPEMIG agencies.

References

[1] B. M. Althouse, J. D. West, C. T. Bergstrom, and T. Bergstrom. Differences in impact factor across fields and over time. *JASIST*, 60(1):27–34, 2009.

[2] S. D. J. Barbosa and C. S. de Souza. INTERACTING WITH PUBLIC POLICY: Are HCI researchers an endangered species in Brazil? *ACM Interactions Magazine*, 18(3):69–71, 2011.

[3] J. Bollen, H. V. de Sompel, A. A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *CoRR*, abs/0902.2183, 2009.

[4] L. Bornmann and H.-D. Daniel. Universality of Citation Distributions - A Validation of Radicchi et al.'s Relative Indicator cf = c/c0 at the Micro Level Using Data From Chemistry. *JASIST*, 60(8):1664–1670, 2009.

[5] L. Bornmann, R. Mutz, and H.-D. Daniel. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *JASIST*, 59(5):830–837, 2008.

[6] Q. L. Burrell. Hirsch's index: A stochastic model. *J. Informetrics*, 1(1):16–25, 2007.

[7] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, chapter 5, pages 127–150. Kluwer Academic Publishers, 2000.

[8] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Procs. of TREC*, Gaithersburg, MD, USA, 2009.

[9] J. Claro and C. A. V. Costa. A made-to-measure indicator for cross-disciplinary bibliometric ranking of researchers performance. *Scientometrics*, 86(1):113–123, 2011.

[10] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonçalves, and A. H. F. Laender. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *JASIST*, 61(9):1853–1870, 2010.

[11] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[12] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender. A brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, 41(2):15–26, 2012.

[13] F. Franceschini and D. A. Maisano. Proposals for evaluating the *regularity* of a scientist's research output. *Scientometrics*, 88(1):279–295, 2011.

[14] V. P. Freire and D. R. Figueiredo. Ranking in collaboration networks using a group based metric. *J. Braz. Comp. Soc.*, 17(4):255–266, 2011.

[15] W. Glänzel and A. Schubert. A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3):357–367, 2003.

[16] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

[17] R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50:361–365, 1996.

[18] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Inf. Process. Manage.*, 36(2):207–227, 2000.

[19] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[20] D. Lee, J. Kang, P. Mitra, C. L. Giles, and B.-W. On. Are your citations clean? *Commun. ACM*, 50(12):33–38, 2007.

[21] L. Leydesdorff. Alternatives to the journal impact factor: I3 and the top-10% (or top-25%?) of the most-highly cited papers. *Scientometrics*, 92(2):355–365, 2012.

[22] J. Lundberg. Lifting the crown - citation z-score. *J. Informetrics*, 1(2):145–154, 2007.

[23] E. A. Oliveira, E. A. Colosimo, D. R. Martelli, I. G. Quirino, M. C. Oliveira, L. S. Lima, A. C. Simões E Silva, and H. Martelli-Júnior. Comparison of Brazilian researchers in clinical medicine: are criteria for ranking well-adjusted? *Scientometrics*, 90(2):429–443, 2012.

[24] I. Podlubny. Comparison of scientific impact expressed by the number of citations in different fields of science. *Scientometrics*, 64(1):95–99, 2005.

[25] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, 105:17268–17272, 2008.

[26] A. Veloso, W. Meira, Jr., M. Gonçalves, and M. Zaki. Multi-label lazy associative classification. In *Procs. of ECML/PKDD*, pages 605–612, 2007.

[27] A. Veloso, W. Meira Jr., and M. J. Zaki. Lazy associative classification. In *Procs. of ICDM*, pages 645–654, Washington, DC, USA, 2006.

[28] L. Waltman, N. J. van Eck, T. N. van Leeuwen, M. S. Visser, and A. F. J. van Raan. Towards a new crown indicator: Some theoretical considerations. *CoRR*, abs/1003.2167, 2010.