Al-Kashaf: Research Journal for Social Sciences (Miscellaneous)
Volume:04, No:04. October–December 2024
ISSN:2790 –8798; E-ISSN: 2790–8801

Correlation of Basic Research Evaluation Indices

***Sohaib Latif, **Muhammad Waleed Butt, ***Haroon Shafique**

***** *Department of Computer Science & Software Engineering, Grand Asian University Sialkot, Pakistan.*

****** *Department of Social Sciences, Grand Asian University Sialkot, Punjab, Pakistan.*

******* *PhD Scholar, University of Gujrat, Punjab, Pakistan.*

KEYWORDS

Citations count
Expert finders
Expertise
g-index
h-index
Research evaluation parameters

ABSTRACT

Research evaluation parameters are a benchmark to measure the scientific output of researchers. Many techniques have been used previously to measure scientific output. Although all these parameters gave a good measure of the researcher's contribution, due to different domains and small volume of data sets, it is hard to say which parameter best measures the expertise of a researcher. This paper analyzes the application of basic research evaluation parameters on a common large dataset in a single domain and investigates their correlation. Firstly, ranking lists of indices were created to analyze the application of parameters. Secondly to investigate their relationship potential correlation of indices was accessed. The research work presented here concentrated on the Computer science domain however we suggest it should apply to other scientific domains as well.

Introduction

Evaluating authors is important in different scenarios because they are needed in all fields of life, especially for taking guidance from them in their relevant subjects, moreover, employers need such high-profile people for doing their projects (Beel, Gipp, Langer, & Breitinger, 2016). We can identify field experts manually and by using automated tools. By using manual techniques, the basic problem is of manual creation of profiles that are not updated regularly (Bornmann, Mutz, & Daniel, 2008). This technique is suitable to identify skilled individuals within the organizations. To counter the problems of manual systems, automated tools for searching competent authors were developed (Buckley & Voorhees, 2017). There are a lot of papers published in different journals and conferences to evaluate researcher's performance (Cole & Eales, 1917). For this purpose, different parameters have been used, such as: publication (Egghe, 2006), citations count (Egghe, 2006), h-Index (Egghe, 2007) and g-index (Fang & Zhai, 2007) etc. Many comparisons have been made by some authors, which show that these parameters have some similarities and variances in their results. In all above parameters, authors have proved their own parameters' efficiency by using different datasets in different domains (Narin, 1976). Due to different datasets, it is hard to say that which parameter best measures the expertise of a researcher.

With the increasing competition among scientific researchers, the need for better indices ranking is also increasing. There are many measuring benchmarks that work well to assess the worth of scientific

*Corresponding Author: Sohaib Latif
Email Address: sohaiblatif095@gmail.com

output. Many authors have evaluated different ranking parameters; however, everyone used a small volume of dataset. Such as: Bornmann et al. (Gross & Gross, 1927) evaluated h-index and its variants by using limited dataset of 414 authors. Meho et al. (Hirsch, 2005) used dataset of 22 researchers to compare citation ranking and h- index. Hirsch, 2007 find the correlation between h- and h(2) indexes and used very limited dataset of 19 professors only. Latif, Fang, Mohsin, Akber, Aslam, Mujlid, & Ullah, 2023 accessed the correlation between h- and g- index and used data set of 168 authors. Further Buckley et al. describes that error rate and number of documents used for measurement have a strong inverse relation to each other. Moreover, authors of different parameters used different datasets on different domains and proved efficiency of their own parameter (Latif, Fang, Arshid, Almuhaimeed, Imran, & Alghamdi, 2023). Motivated from above findings we thought to evaluate different indices on a common large dataset and investigates the correlation between these indices in a single comprehensive domain of Computer Science.

Methodology

The architecture diagram is shown in Figure 1 which shows all steps to carry research from data collection to correlation. Whole research work is divided in four phases. The first phase was the data collection phase. For this purpose, we used ACM classification to collect Meta data, because it is a certified classification in Computer Science domain. Moreover, many popular journals of science such as: JACM (Journal of ACM), CSUR (Computing Surveys), JEA (Journal of Experimental Algorithmic), etc. also use ACM classification. But the problem was; that there were many categories present in this which were too broad that getting relevant data on them was impossible. For example, there was a keyword ―fan‖ that may point to some other thing or features. For this reason, we contacted the domain experts and compiled a modified list of categories that gave us relevant data.

Second phase was preprocessing. In this phase we manually checked and refined data where it required. The data refining phase explore data in two dimensions: in the first dimension we checked author disambiguation and in second dimension we checked paper relevancy. Author Disambiguation was one of the major issues in Google scholar. This issue arose when common name occurred for multiple authors; we match their first names and then first and last names. Especially the issue was in last name of authors, last name of some authors was common. We checked manually and removed duplications.

To verify dataset in second dimension, many concentrate steps were taken. For example, (i) Invalid characters in their title were removed. (ii) The Journals/conferences were verified. Some papers have title names with some character like @, #, &, *. We removed such characters from their titles. Some papers were found irrelevant that does not belong to Computer Science domain; we also removed such papers. After removing irregularities, we stored purified dataset in our database. In the process of data refining 32607 papers were sorted out of 32821 papers and 214 papers were discarded because of author ambiguity and irrelevancy.

Third phase was indices evaluation where we applied whole common dataset of 55556 authors on udder study parameters (Publication count, Citations count, h-index, and g-index).to investigate the application in

computer science domain. We used programming language CSharp and Visual Basic for evaluation purpose and calculated the author ranking of research evaluation parameters under study.

The Fourth phase presents the correlation evaluation in which we found correlation of research evaluation indices. Correlation is the extent to which two or more values/variables relate or differ with each other. We used M.S Excel, an automated tool, to find correlation of parameters (publication count, citations; h index and g- index) with one another. The main purpose of correlation was to access similarities and variances between indices and also to know the behavior of these indexes towards each other.

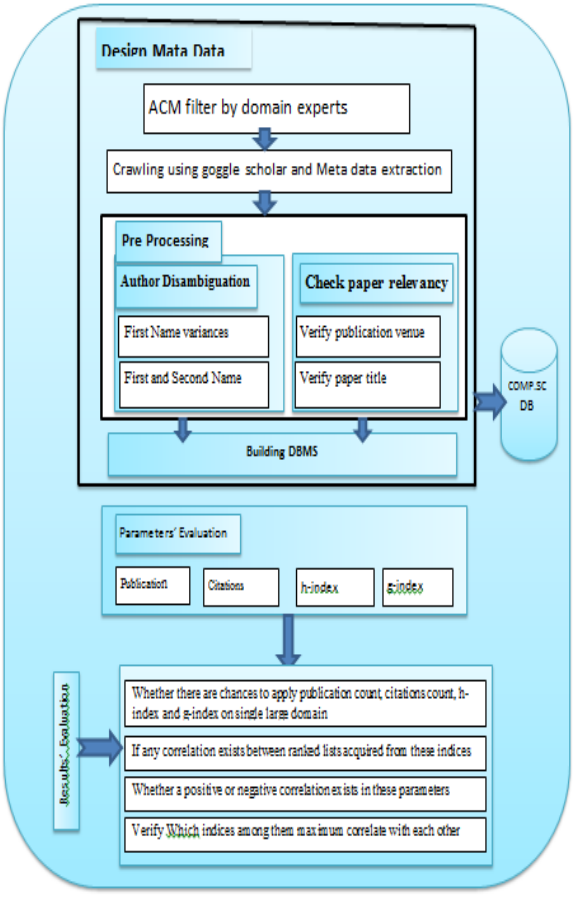


Figure 1: Architecture Diagram

Definitions of Indexes

This section of the paper gives brief introduction of the evaluation parameters under study.

A. Publication count

The earliest Bibliometric paper for research evaluation was produced by Cole and Eales in 1917. In which they analyzed the literature of Anatomy from the year 1543 to 1860. Their aim was to measure the performance of participating countries during past three centuries. Publication count is one of the commonly used parameters in which only the total numbers of publications by the author are counted to judge knowledge in scientific domain and measure the skill of researchers in scientific domain. The author who has number of publications more than others is considered to be the most prestigious researcher and is placed at the top in the ranking list. Publication count is a simple parameter that can be calculated only by counting total no. of publications of the researchers' research unit from his first publication. Publication count is a quantitative benchmark that measures overall productivity of an author, but the main flaw of publication count parameter is that publication count alone may not reflect the overall knowledge of an individual and it does not measure impact of scientific research.

B. Citations Count

In 1955 Garfield introduced a new benchmark that can measure the overall impact and quality of the work. The concept of the citations count was first presented by Gross and Gross (Yimam-Seid & Kobsa, 2003) to measure the adequacy of the college library. However, the citation as a measure was introduced by Garfield in 1955. Citation

is a quantitative parameter to measure the researcher's ranking based on the total number of citations that a researcher has received for his papers. Most commonly a researcher is most prestigious if his/her papers are cited frequently by the other researchers.

―Citation analysis consists of counting citations to publications of a researcher's research unit, then comparing citations with the Citations of other researcher' research unit of similar documents (Ravichandra Rao, 2007).

It is a good measure of total impact of papers however, may be changed by a small number of ‖ big hits ‖, which may not be representative of the researcher if he/she is co-author of the others on those papers. Moreover, citations count gives undue weight to highly cited review articles versus original research contributions (Taylor & Richards, 2008).

C. h-index

To overcome the deficiencies of publication count and citation count parameters h-index was introduced by Jorge E Hirsch in 2005 as an evaluator to measure the scientific research output. Hirsch defined the h-index as:

―A scientist has index h if h of his/her Np papers have at least h citations each and other (Np – h) papers have < h citations each,

It combines two types of measures (publication count and citation count), so has more benefit over such single numbers. h-index is a good predictor as compared to other parameter while we predict future achievements. In determining the h-index highly cited papers are important, but once they belong to authors' h-index, no matter it receives more citations or not.

D. g-index

After the publication of the h-index to cover its flaws a new index was presented by Egghe. It is the modification of the h-index. g-index is defined as:

― The highest number g of papers that received g^2 or more citations.‖ (Hirsch, 2007).

g-index eliminates the limitations of h-index that a paper once belongs to top h papers then the subsequent citations are not counted [9]. Moreover g-index is not limited to publication period instead it covers the period from scientists first publication to ranking date. However, g- index has also some flaws: if a researcher receives a small number of ―big hits‖ (a few papers receives much citations) then his g-index will increase a lot as compared to other researchers, who have average citations count.

Results and Evaluation

The following section presents the analysis of the ranking of research evaluation parameters on a larger dataset in the Computer science domain. A total of 32607 papers that belong to 55556 different authors of various research fields of computer science were used in the evaluation. We used C-sharp and Visual Basic for evaluation.

Publication Count

Publication count is the original production of an author. Figure 2 is a graph that shows the publication count of each author against his author_ id. Our publication count ranking proved as obvious from the graph that a maximum number of publications in our

dataset is 88. Publication count is the simplest way to rank the authors, the more publications an author has, the high will be his publication count. Moreover, the rank of an author is directly proportional to the number of publications of the author, so the publication count of researchers increases with the increase in the research unit of the researcher.

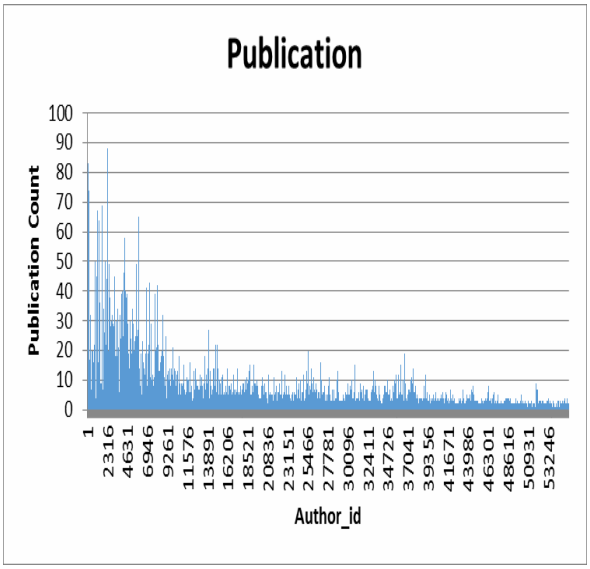
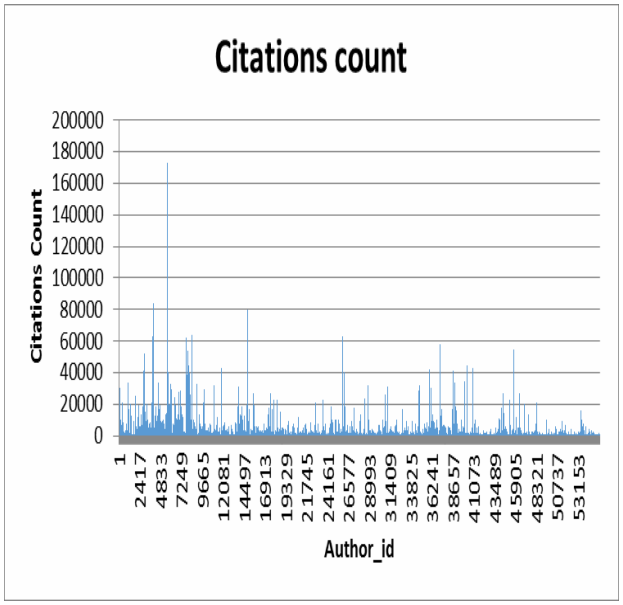


Figure 2: Publication Index ranking using refined data

Citations Count

Citations count parameter ranks the authors as per the commutative sum of the citations of his publications. The figure 3 below is a graph that represents the Citations count of each author against his author_ id. Citations count ranking list proved as also the graph shows that maximum Citations count in our dataset is 173186 and minimum Citations count is zero. Out of 55556 authors in our dataset 55016 authors got at least one citation each, but 540 authors failed to get any citation. So, they have rank zero.

Figure 3: Citations count ranking using refined data



This is because that citations rank depends not only on the number of papers but also accounts for the number of citations that each paper received.

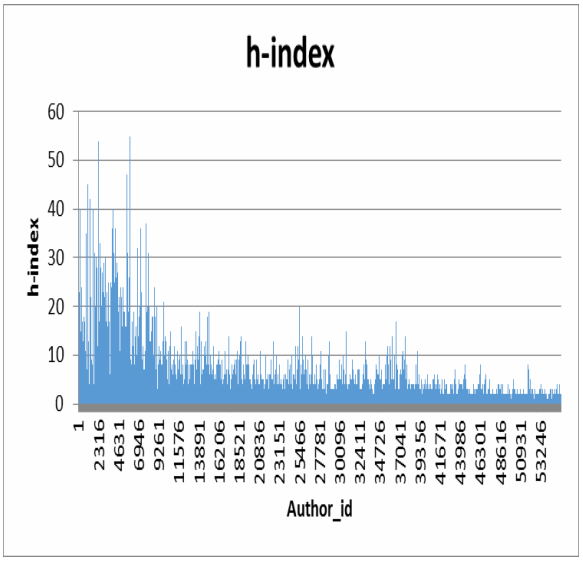
h-index

—A scientist has index h if h of his/her N_p papers have at least h citations each and other (N_p – h) papers have < h citations each.

In below figure 4 the graph shows h-index of each author against his author_ id. H-index ranking list proved as also the graph represents that maximum h-index rank in our dataset is 55 and minimum h-index is zero. Out of 55556 authors in our dataset 99.03% got at least one h-index each, and only 0.97% were still with zero h-index value. The value of h-index does not increase with the increase in number of publication or citations alone, instead to increase value of h, number of publication and citations must increase accordingly.

Moreover, the increment of the publications only or citations received for

Figure 4: h-index ranking using refined data



papers already included in h-index does not affect the ranking of the author.

g-index

G-index is defined as: ―The highest number g of papers that received g^2 or more citations ‖.

The graph below in Figure 5 represents the value of the g-index for each author against his author_ id. g-index ranking list proved as also depicted in the graph that the maximum value of the g-index in our dataset is 88 and the minimum g-index is zero. Out of 55556 authors, 540 authors were with zero g-index values because they were unable to get any citations. The value of the g-index is always equal to or greater than the h-index because it may increase by increasing the citations received by the papers that have already been included in the index ranking while the h-index does not.

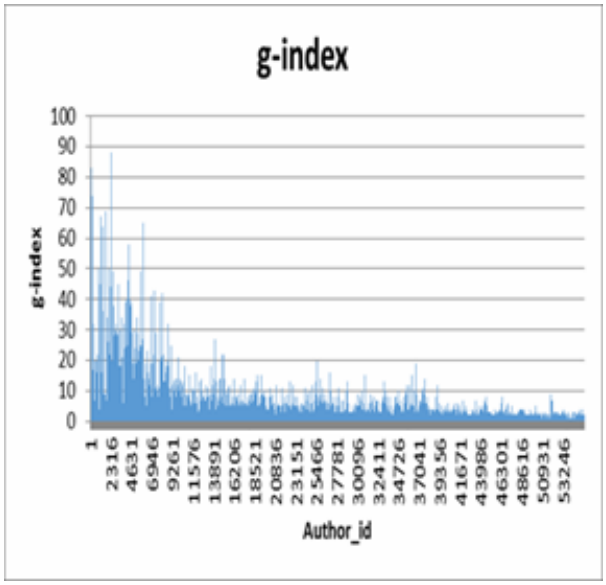


Figure 5: g-index ranking using refined data

The following section presents a detailed analysis of the correlation between research evaluation parameters on a large dataset in the computer science domain. The nature of the correlation of indices is also discussed in this section. We used an automated feature of M.S Excel to find correlations,

Correlation Analysis

To evaluate the correlation, we took ranked lists of indices (Publication count, Citations count, g-index, and h-index) generated earlier in the parameters' evaluation and accessed their correlation using M. S Excel.

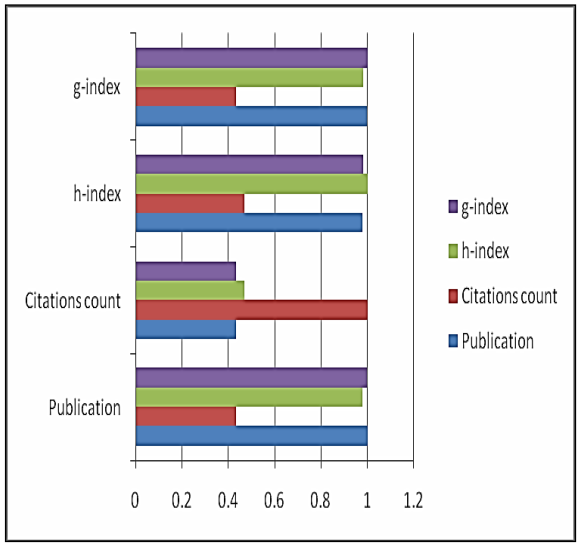


Figure 6: Correlation analysis of all indices

We found the collective correlation of all (four) parameters to analyze the overall correlation of our parameters. The graph depicted in Figure 6 above represents that correlation values of indices are as under:

i. Publication Count vs. Citations Count = 0.435083
ii. Publication Count vs. h-index = 0.979032
iii. Publication Count vs. g-index = 0.998798
iv. Citations Count vs. h-index = 0.469530
v. Citations Count vs. g-index = 0.435075
vi. h-index vs. g-index = 0.990858

The correlation analysis evidences that Publication count and g-index are maximally correlated with each other, with a high correlation value of 0.998798, while citation count and g-index (correlation value = 0.435075) are weakly correlated, with only 43% similarity between them. All other indices have a correlation (with each other) between these two peak values, as also presented below in Table 1.

Table 1 Correlation analysis of publication count, citation count, and indices

	Publication	***Citations count***	***h-index***	***g-index***
Publication	***1***	***0.435083***	***0.979032***	***0.998798***
Citations count	***0.435083***	***1***	***0.469530***	***0.435075***
h-index	***0.979032***	***0.469530***	***1***	***0.980858***
g-index	***0.998798***	***0.435075***	***0.980858***	***1***

Conclusion and Future Work

The study results reveal that applying different evaluation parameters on a common large dataset of a single domain is equally possible. The analysis of the evaluation parameters shows that all the parameters under study (except publications count) depend mainly on the citation count of the author's paper. That is why authors with zero citation count have zero values for h-index and g-index as well. Further correlation analysis evidenced that a positive correlation exists between publications count, citation count h-index, and g-index. However, among all, publication count, and g-index have a maximum correlation between them, so they have a strong association with each other whereas citation count and g-index have the low association to each other.

Our future intentions include applying other evaluation parameters such as variances of h-index on the computer science domain and identifying correlations between them. We will also explore the effect of under-study parameters on the data of other domains and validate if the same findings apply to the data of other domains. We are interested in revealing the factors affecting the correlation of indices. In the future we are also planning to predict new evaluation parameters based on indices' correlation which will be an efficient measure of

scientific research output as well as good a predictor of the future achievements of researchers.

.

References

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: A literature survey. International Journal on Digital Libraries, 17(4), 305–338.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). 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. Journal of the American Society for Information Science and Technology, 59(5), 830–837.

Buckley, C., & Voorhees, E. M. (2017). Evaluating evaluation measure stability. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 235–242). ACM.

Cole, F. J., & Eales, N. B. (1917). The history of comparative anatomy: Part I.—A statistical analysis of the literature. Science Progress (1916-1919, 11(44), 578–596.

Egghe, L. (2006). An improvement of the H-index: The G-index. ISSI Newsletter, 2(1), 8–9.

Egghe, L. (2006). Theory and practice of the g-index. Scientometrics, 69(1), 131–152.

Egghe, L. (2007). Dynamic h-index: The Hirsch index in function of time. Journal of the American Society for Information Science and Technology, 58(3), 452–454.

Fang, & Zhai, C. (2007). Probabilistic models for expert finding. Lecture Notes in Computer Science: Advances in Information Retrieval (pp. 418–430).

Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. Science, 66(1713), 385–389.

Hirsch, J. E. (2005). Expertise Browser: An index to quantify an individual's scientific research output. International Conference on Software Engineering (pp. 503–512).

Hirsch, J. E. (2007). Does the h-index have predictive power? Proceedings of the National Academy of Sciences, 104(49), 19193–19198.

Kelly, C. D., & Jennions, M. D. (2006). The h-index and career assessment by numbers. Trends in Ecology & Evolution, 21(4), 167–170.

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. ISSI Newsletter, 2(3), 4–6.

Latif, S., Fang, X., Mohsin, S. M., Akber, S. M. A., Aslam, S., Mujlid, H., & Ullah, K. (2023). An enhanced virtual cord protocol-based multicasting strategy for the effective and efficient management of mobile ad hoc networks. Computers, 12(1), 21.

Latif, S., Fang, X. W., Arshid, K., Almuhaimeed, A., Imran, A., & Alghamdi, M. (2023). Analysis of birth data using ensemble modeling techniques. Applied Artificial Intelligence, 37(1), 2158273.

Narin, F. (1976). Evaluative bibliometric: The use of publication and citation analysis in the evaluation of scientific activity. Washington, DC: Computer Horizons.

Ravichandra Rao, I. K. (2007). Distributions of Hirsch-index and G-index: An empirical study. In D. Torres-Salinas & H. F. Moed (Eds.), Proceedings of the 11th Conference of the International Society for Scientometrics and Informetrics (Vol. 2, pp. 655–658).

Taylor, M., & Richards, D. (2008). Discovering areas of expertise from publication data. In Pacific Rim Knowledge Acquisition Workshop (pp. 218–230). Berlin, Heidelberg: Springer.

Yimam-Seid, D., & Kobsa, A. (2003). Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach. Journal of Organizational Computing and Electronic Commerce, 13(1), 1–24.