Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Analysis of H-index and Papers Citation in Computer Science Field using K-Means Clustering Algorithm

Omar Ibrahim Obaid[1*]

[1] Department of Computer, College of Education, Al-Iraqia University, Baghdad, Iraq

*Corresponding Author: Omar Ibrahim Obaid

DOI: https://doi.org/10.52866/ijcsm.2023.02.02.006
Received october 2022; Accepted janeuary 2023; Available online February 2023

ABSTRACT: This paper provides an analysis of H-index and paper citations in the computer science field using K-Means Clustering Algorithm. By leveraging cutting edge visual analytics through the use of Power BI and Orange Data Mining tool with K-Means clustering algorithm, we are able to present a comprehensive review of how H-index and citations impact the scholarly evaluation of authors in computer science field. The analysis obtained will assist academics seeking to expand their research influence while providing additional context to the general audience seeking to understand how machine learning algorithms can contribute to data exploration. The analysis conducted has revealed that area of research which has the highest value of citation and H-index include the following topics: Artificial-Intelligence, Computational-Intelligence, Data-mining, Evolutionary-Algorithms, and Big Data Analytics. Finally, the analysis of this research clearly demonstrated that paper citations remain an important factor for determining scientific impact in disciplines such as computer science given its close relationship with established metrics such as h-index scores.

Keywords: H-index, Papers Citation, Computer Science, K-means Clustering, Power BI, Analysis

1. INTRODUCTION

The H-index is a metric used to measure the productivity and impact of an academic paper. It was first proposed by (Jorge E Hirsch) in 2005 and was defined as "the maximum number of papers, each having received at least 'H' citations". The most commonly accepted definition of the H-index is that it is ‘the largest number of papers with citation count equal to or higher than that number’. In other words, if an author has published ten papers and each paper has been cited at least ten times, then the author would have an H-index of 10 [1,2].

Furthermore, Paper citations are important in measuring the impact of an academic article because they show how often other authors are citing a particular work. Citation counts provide evidence for the relevance and importance of a publication within the research community. Citation metrics have been around since 1904 when Jean Perrin introduced them into scientific discourse, and for centuries scholars have used this type of metrics to assess scholarly influence and evaluate researcher performance [2, 3].

In general, high numbers of citations indicate that a paper is influential or significant within its field. Journals with higher rejection rates also tend to publish more highly cited papers as they usually favor highly regarded works over those with lower citation counts. Moreover, using paper citation analysis can help researchers determine what topics are being discussed most frequently and which papers are having the greatest impact on their fields [4] .

Paper citations are often utilized to assess the quality, significance, and importance of a research paper in the academic field. On the other hand, in software development, success is measured by the quantity of downloads. Unfortunately, only a few papers reach high levels of citation while many tend to go unnoticed. Overall, research has shown that certain characteristics like paper quality, journal prestige, number of authors, accessibility level and

*Corresponding author: omar.obaid@aliraqia.edu.iq

collaborating parties play a bigger role than other factors like researcher statistics like race, age or gender when it comes to influencing citation rate [5].

The aim of this paper is to determining widely cited research fields, recognizing scientists of computer science with high h-index rankings, and identify various patterns existing among different authors that can be useful for further research studies for computer science field. In conclusion, while there are numerous ways in which scientists can be evaluated, using metrics such as the h-index and paper citations allows researchers to better assess their own work as well as that of their peers. This can be helpful in understanding the current state and future directions in any discipline or research area.

2. LITERATURE REVIEW

Studies have demonstrated that the higher one's h-index, the greater their impact on academic literature was perceived to be. Similarly, studies found that citation analysis alone can successfully distinguish elite scholars from non-elite ones based on their published works [6, 7]. It was also discovered that there were strong correlations between total number of citations and the h-index, which indicated that authors with large numbers of citations are likely to have a high h-index as well. Furthermore, research has shown that when considering an author's work, it is beneficial to use both metrics in tandem rather than relying on just one measure [8].

In addition to uncovering insight concerning individual scholars' standing in academia, relative measures such as correlation between H-index and paper citation have also served as broader indicators for scientific fields. Analysis has shown, for instance, that there existed significant differences in mean values between disciplines regarding their H-indices distribution [8, 9]. Thus, comparative cross discipline analyses led by the combination of H-Index and papers citation frequency appears to be a viable approach for assessing scholarly productivity and influence across various domains.

Overall, h-Indexes combined with paper citations present valuable tools for studying both an individual scholar's influence as well as factors determining scientific prestige. Both these metrics offer researchers comprehensive data points from which decisions about authorship or article adherence can be made along with broad overviews concerning certain subfields within science at large.

3. MATERIALS AND METHODS

3.1 MATERIALS

The analysis in this paper utilized a dataset which is downloaded from Kaggle and the CSV file is obtainable from Ref [10]. It contained data from 2016–2020 and about 103 researchers associated with Computer Science field including their names; affiliation; total papers citation counts; h-index, citation of 2020; citation since 2016; and h-index since 2016.

This dataset is clean and has no missing value and ready for analysis. Furthermore, we have used Microsoft Power BI to make the analysis of this study. Power BI is a business analytics service provided by Microsoft that provides data visualization and interactive functionality as an easy way to visualize and analyze various types of data [11].

On the other hand, Orange Data Mining is used to implement K-Means clustering algorithm. Orange Data Mining is a powerful open-source data mining tool that enables users to explore and analyze data using a variety of graphical tools. It supports interactive data visualization, text mining and machine learning [12]. The following table shows the statistics of the dataset used in this paper which was obtained using Power BI.

Table 1. Statistics of the Dataset

Features	Mean	Median	Dispersion	Min.	Max.	Missing
H-. Index	75.5481	67.5	0.437813	2	157	0
Citation 2020	4484.71	3575	0.637834	14	12012	0
Total Citation	31649.4	20735.5	0.888112	27	131149	0
H-Index Since2016	64.2115	61.5	0.377982	2	123	0
Citation Since	21358.4	15992.5	0.698947	27	65762	0

2016

3.2 METHODS

In this paper, we used K-Means clustering which is a popular and widely used unsupervised machine learning algorithm for data clustering. It is an iterative process that partitions the given data into k clusters, where each cluster has its own centroid.

The goal of the K-Means algorithm is to minimize the within-cluster variance, which means that all points in a single cluster should be as close as possible to its centroid or mean point. The K-means algorithm works by randomly initializing k centroids then assigning each observation in our dataset to one of these clusters based on their Euclidean distance from the centroids. This continues until no observations change clusters or when some other stopping criteria have been met such as maximum number of iterations has been reached. After this step, we can calculate new centers for our newly formed clusters and repeat this procedure until convergence occurs (all points belong to same cluster) or minimum within group sum of squares criterion value achieved [13].

In conclusion, K Means Clustering Algorithm provides an effective way for grouping datasets with similar characteristics together without any prior knowledge about them; making it suitable for exploratory analysis tasks due its simplicity and flexibility.

4. RESULTS AND DISSCUSION

4.1 POWR BI ANALYSIS

The analysis that was conducted yielded several conclusions about the determining widely cited research fields, recognizing scientists of computer science with high h-index rankings, and identify various patterns existing among different authors that can be useful for further research studies for computer science field. The following figure shows the average of H-Index by citation since 2016.

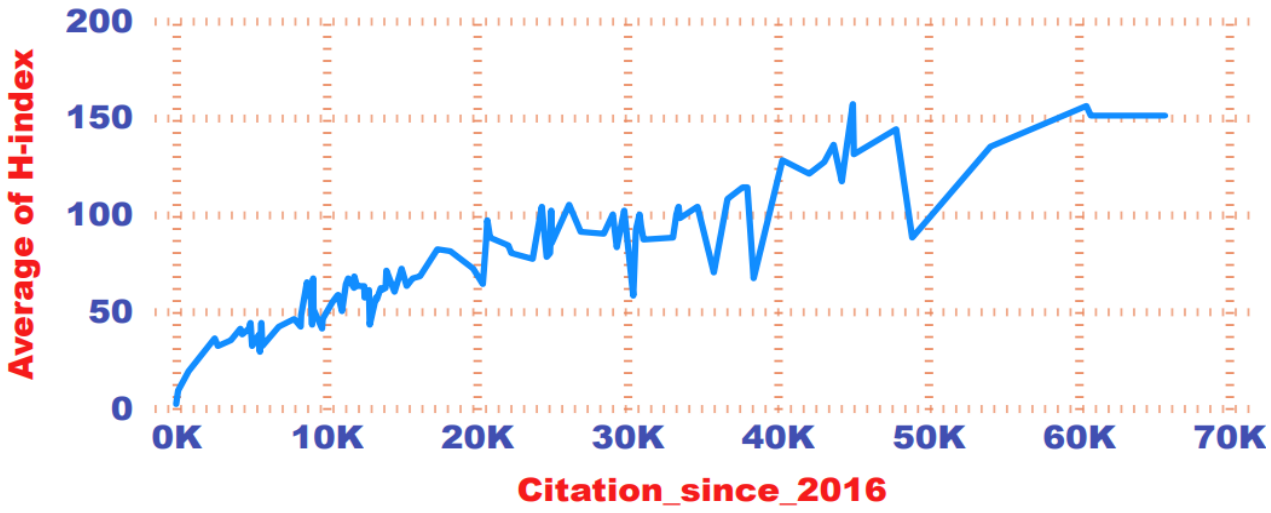

FIGURE 1. Average of H-Index by citation since 2016

The data also revealed interesting patterns when examining authors' performance with respect to their total paper citations and respective h-index values. In general, those authors who tended to be more highly cited across a wider variety of topics had higher h-index values than authors whose papers tended to be narrowly focused on one topic area only. Figure 2 illustrates the average of H-Index by citation since 2020.

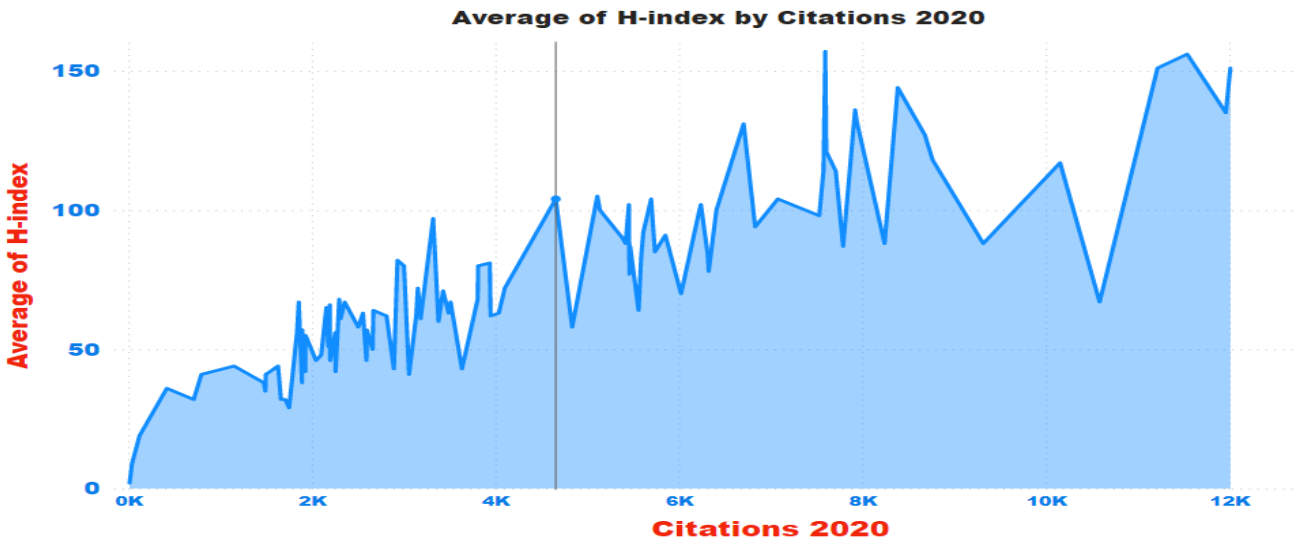

FIGURE 2. Average of H-Index by citation since 2020

Moreover, table 2 illustrates the scientists with higher H-Index, while table 3 shows the sample of research area that received the highest H-Index. Finally, table 4 demonstrates the sample of research area that received the highest citation.

Table 2. The scientists with higher H-Index

Name	H-index
Dr. P. N. Suganthan	100
Victor C. M. Leung	100
Geoffrey Ye Li (李烨)	102
Timon Rabczuk	102
Jeffrey Andrews	104
L Hanzo	104
Qing-Long Han	104
Enrique Herrera-Viedma	105
U Rajendra Acharya	108
Xuelong Li	114

Table 3. Sample of research area with highest H-Index

H-Index	Area of Research
156	Artificial Intelligence / Computational Intelligence/ Data mining/ Evolutionary
151	Computer Systems/ Cloud Computing/ Distributed System/ Edge Computing / Internet of Things
151	Information Theory/ Statistical Signal Processing/ Stochastic Analysis
144	Systems and Control/ Electrical and Electronic Engineering/ /cybernetics/ Human-Computer Interaction
136	Decision Making/ Information fusion/ Fuzzy Sets/ Computational Intelligence/ Bibliometrics
135	Artificial Intelligence/ Machine Learning/ Computer Vision/ Image Processing/ Data Mining
132	Wireless/MIMO/Communication/Signal Processing/Information Theory
131	Wireless Communication/ 5G/6G Wireless System/ Intelligent Surfaces for Wireless Communication/Nano-Scale and Molecular Communication
128	Complex Network/Neural Network/Multi- Agent System-Engineering Stability-Dynamics and Control- Time-Delay- Systems
127	Wireless Networking/ MAC/Network Security/VANET
121	Cybersecurity/IOTs/Mobile Nets/Big Data Analytics/Cloud Computing/ Artificial Intelligence/Machine Learning

Table 4. Sample of research area with highest total citation

H-Index	Area of Research
131149	Wireless Communication/ 5G/6G Wireless System/ Intelligent Surfaces for Wireless Communication/Nano-Scale and Molecular Communication
119215	Computer Systems/ Cloud Computing/ Distributed System/ Edge Computing / Internet of Things
115279	Information Theory/ Statistical Signal Processing/ Stochastic Analysis
105322	Artificial Intelligence / Computational Intelligence/ Data mining/ Evolutionary Algorithms/Big Data Analytics
77545	Wireless/MIMO/Communications/Signal Processing/Information Theory
73276	Systems and Control/ Electrical and Electronic Engineering/ /cybernetics/ Human-Computer Interaction
70084	Artificial Intelligence/ Machine Learning/ Computer Vision/ Image Processing/ Data Mining
70021	Decision Making/ Information fusion/ Fuzzy Sets/ Computational Intelligence/ Bibliometrics

68885	Wireless Networking/ MAC/Network Security/VANET
66102	Complex Network/Neural Network/Multi- Agent System-Engineering Stability-Dynamics and Control- Time-Delay- Systems
64401	Swarm Intelligence/Optimization/ Metabeuristics/Computational Intelligence
64221	Wireless Communications/Communication Theory/ Machine Learning for Communications/Stochastic Geometry/Information Theory
63036	Telecommunications/Wireless/ Communication/Quantum-Information-Science/Video-Coding

4.2 K-MEANS CLUSTERING ANALYSIS

Regarding K-Means clustering algorithm, the results showed that total paper citations and H-Index were heavily concentrated within 4 clusters, with cluster 2 being the largest, followed by clusters 3. Among these clusters, the highest average h-index and total citations value were observed in cluster 4, indicating that this cluster had the greatest influence on scholarly impact across all fields. The research area of this cluster which has the highest citation and h-index is demonstrated in table 3 and 4 respectively. The following figure illustrates the clusters with H-Index by total citations.

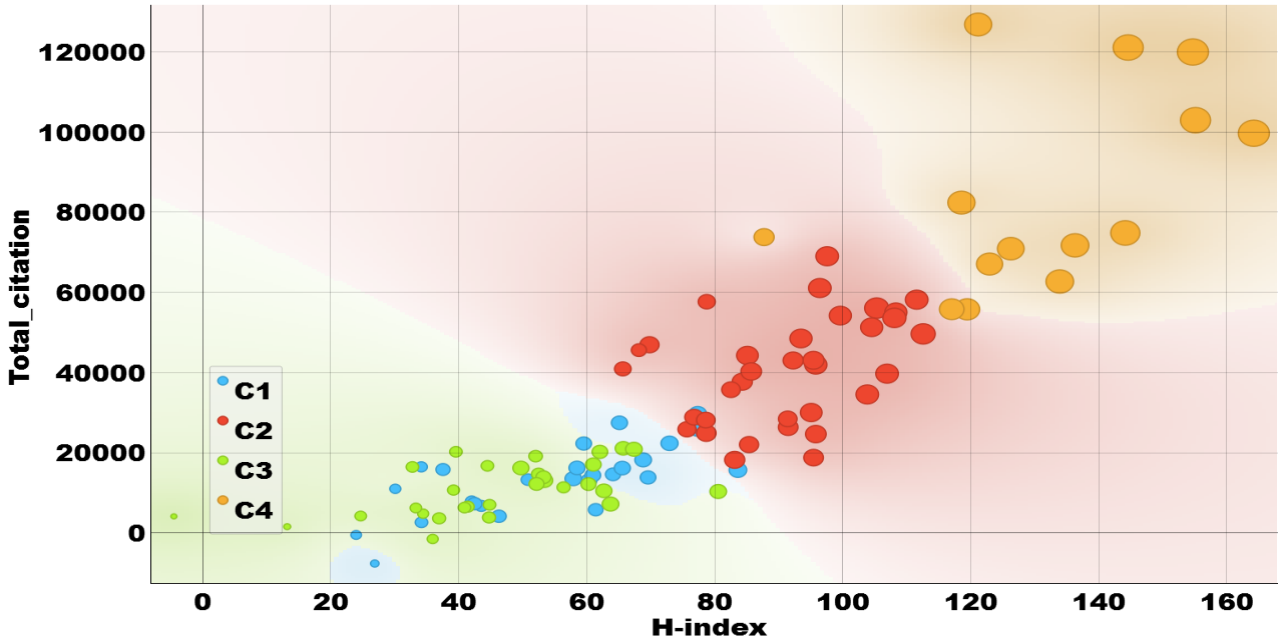

FIGURE 3. The clusters with H-Index by total citations

Furthermore, Silhouette analysis is used to study the separation of clusters in a dataset. Generally, it is used for determining the validity of a clustering algorithm. The silhouette coefficient for each data point can be interpreted as follows:

- A value of 1 indicates a perfect match between the data point and its cluster .
- Values between 0 and 1 indicate that the data point is closer to its own cluster than to other clusters .
- Values lower than 0 indicate that the data point may have been assigned to the wrong clustering

One can interpret a silhouette analysis figure by looking at both the average silhouette scores as well as individual silhouette scores for each point. High values indicate that partitions created by the clustering algorithm are well separated, and thus valid; low average scores suggest otherwise. Additionally, one can also look at individual points' silhouette scores to identify any outliers or misclassified points due to noise or improper parametrization of the clustering algorithm itself. The following figure shows the silhouette analysis of the clusters.

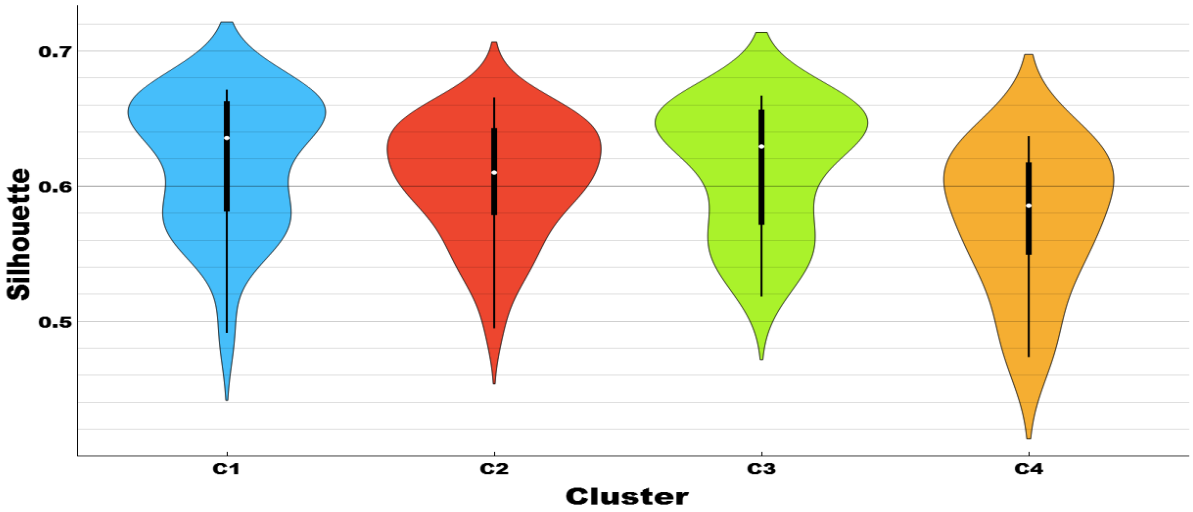

FIGURE 4. The silhouette analysis of the clusters.

In addition, application of the k-means clustering algorithm divided paper citation into four groupings based on similarity of citation patterns among researchers from different subfields of computer science. Furthermore, the following figure demonstrates the silhouette analysis of H-Index and total citation with regression line.

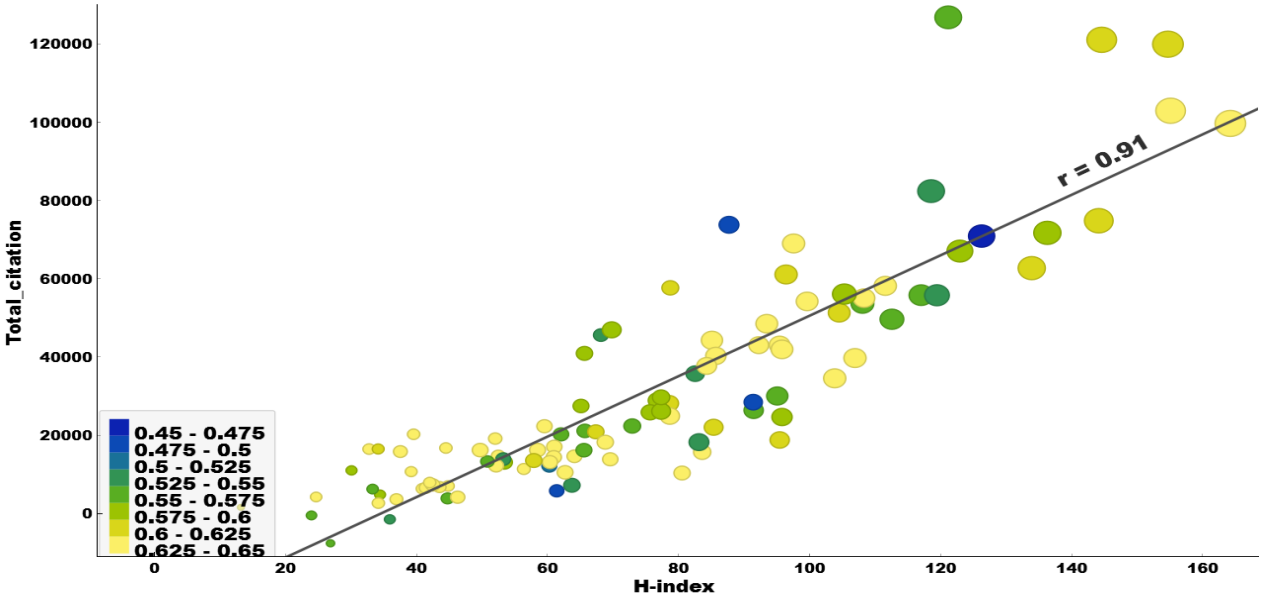

FIGURE 5. The silhouette analysis of H-Index and total citation with regression line

Finally, results suggest clusters based on papers citations primarily related to software engineering or information technology have relatively lower overall influence than those clusters related to areas such as artificial intelligence or mathematical methods for computing systems applications. This finding further confirms the assumption that concentrations within select subdisciplines contribute positively toward higher academic impact on scholarly output when evaluated using meaningful metrics like the h-index.

5. CONCLUSION

The use of K-Means Clustering Algorithm along with Power BI for the analysis of H-Index and paper citations in the computer science field have proven to be an effective method for measuring and identifying research performance. Their ability to cluster data by similar characteristics is useful for obtaining more customized and accurate results. This can allow researchers to have a better understanding of their output, as well as providing greater insight into the overall scientific field. Furthermore, K-Means clustering enables much more targeted research since clusters can be used to assess cohesion of network connections or compare the relative success of individual papers in many different ways. As seen from the case study, this combination provides a way for researchers to analyze citation trends for better understanding among their peers.

Funding

None

ACKNOLEDGEMENT

The author we would like to thank the reviewers for their valuable contribution in the publication of this paper.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] H-index (2023) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/H-index (Accessed: February 4, 2023).
[2] Hirsch, J. E. Does the H index have predictive power? Proceedings of the National Academy of Sciences of the United States of America, 104(49), 19193– 19198, 2007.
[3] X. Bai, F. Zhang, and I. Lee, “Predicting the citations of scholarly paper,” Journal of Informatics, vol. 13, no. 1, pp. 407–418, 2019.
[4] T. Yu, G. Yu, P. Li, and L. Wang, “Citation impact prediction for scientific papers using stepwise regression analysis,” Scient metrics, vol. 101, no. 2, pp. 1233–1252, 2014.
[5] C. Castillo, D. Donato, and A. Gionis, “Estimating number of citations using author reputation,” in International Symposium on String Processing and Information Retrieval, N. Ziviani and R. Baeza-Yates, Eds., vol. 4726 of Lecture Notes in Computer Science, pp. 107–117, Springer, Berlin Heidelberg, 2007.
[6] Bai, X., Xia, F., Lee, I., Zhang, J., & Ning, Z. Identifying Anomalous Citations for Objective Evaluation of Scholarly Article Impact. PloS One, 11(9), e0162364, 2016.
[7] V. Jai Kumar , Y.V.Bhaskar Reddy , V.S.Nishok , Vivek K., Customer Data Processing in Internet of Things Environments Using RFM Analysis and K-means Clustering, Journal of Intelligent Systems and Internet of Things, Vol. 7 , No. 1 , (2022) : 08-19 (Doi : https://doi.org/10.54216/JISIoT.070101)
[8] Robbert van Haselen, the h-index: A new way of assessing the scientific impact of individual CAM authors, Complementary Therapies in Medicine, Volume 15, Issue 4, Pages 225-227, ISSN 0965-2299, https://doi.org/10.1016/j.ctim, 2007.
[9] Acuna, D. E., Allesina, S., & Kording, K. P. Future impact: Predicting scientific success. Nature, 489(7415), 201–202. ISSN 0028-0836, 2012.
[10] Devastator, T. Papers citations vs h-index, Kaggle. Available at: https://www.kaggle.com/datasets/thedevastator/cs-researchers-h-index-and-citation-analysis (Accessed: February 4, 2023).
[11] Raheela zaib, & OURLIS Ourabah. (2023). Large Scale Data Using K-Means. Mesopotamian Journal of Big Data, 2023, 38–47. https://doi.org/10.58496/MJBD/2023/006
[12] Bioinformatics Laboratory, Data Mining, Orange Data Mining - Data Mining. Available at: https://orangedatamining.com/ (Accessed: February 4, 2023).
[13] Youguo Li, Haiyan Wu, A Clustering Method Based on K-Means Algorithm, Physics Procedia, Volume 25, Pages 1104-1109, ISSN 1875-3892, https://doi.org/10.1016/j.phpro.2012.03.206, 2012.