EasyChair Preprint

№ 11578

H-Index Analysis of Research Paper Using Web Crawling Techniques

Omprakash Kambli, Aarti Karande and Harshil Kanakia

December 19, 2023

H-Index Analysis of Research Paper using Web Crawling techniques

Omprakash Kambli[1[0009-0001-0103-4670]] , Dr. Aarti Karande[2[0000-0001-9560-1113]],

Prof. Harshil Kanakia[3[0009-0009-1221-1212]]

[1&2&3] Bharatiya Vidya Bhavan's Sardar Patel Institute of Technology (S.P.I.T.) Bhavan's Campus, Munshi Nagar, Andheri (W), Mumbai-58, Maharashtra, India

[1] naikaumprakash@gmail.com
[2] aartimkarande@spit.ac.in
[3] harshil_kanakia@spit.ac.in

Abstract. The evaluation of research impact of publications is a crucial aspect of academic research. In the recent years, h-index has become a widely used metric for assessing research productivity and impact. In this paper, we shall explore the use of various web crawling techniques and Python programming language to collect publication data from Google Scholar and calculate the h-index of academic researchers. We would also demonstrate how the Scholarly package in Python can be used to retrieve publication data and perform h-index calculations, providing an efficient and objective means of evaluating research impact. Our proposed methodology combines the power of web crawling, h-index, and Python to enable comprehensive research analysis and evaluation. The Workflow for various steps of the process have been listed along with their individual steps and the corresponding code which implements the said functionality. Web crawling stands as an integral part of this process since the data is fetched from sources such as google scholar. The data used for analysis has been collected using the code snippets provided. This paper presents a valuable contribution to the academic research community by providing an objective and efficient method for evaluating research impact.

Keywords: Web Crawling, H-index, Python, Research Impact, Google Scholar Metrics.

1 Introduction

The H-index is a bibliometric indicator that has become widely used for examining research performance and impact of scholars. It measures both the quantity and quality of publications that a researcher has done, based on the number of publications and the number of citations those publications receive.

A research paper is a comprehensive written report that presents the findings of a research study or investigation. It is an essential medium for communicating research outcomes to the scientific community and the general public. Research papers can take various forms, such as empirical studies, literature reviews, case studies, and theoretical or conceptual papers. These papers are critical for advancing knowledge in a particular field and driving scientific progress. [1]

We chose to analyze the H-Index of research papers using web crawling techniques because it is widely used in the research domain and is crucial for evaluating the impact of academic publications. We desire to evaluate H-Index as a metric for measuring research significance, enhancing our understanding of research impact assessment in academia.

The H-index provides a measure of the productivity and citation impact of a researcher's work, allowing for a comprehensive assessment of their research contributions. However, the H-index has limitations, such as its inability to distinguish between different types of publications and its lack of sensitivity to the recency and impact of citations. Despite these limitations, the H-index remains a valuable tool for evaluating research impact and is widely used in academic settings. [2]

Recent research has highlighted the strengths and limitations of the H-index as a tool for evaluating research performance. While the H-index is a straightforward and objective measure that can quickly assess the research productivity and impact of scholars, it is influenced by disciplinary coverage and citation practices, which may limit its accuracy in capturing the true impact of research [2]. Additionally, the H-index does not distinguish between different types of publications or account for the recency and impact of citations, which can lead to biased assessments of research performance [3].

Despite these limitations, the H-index remains a widely used and influential bibliometric indicator in research evaluation. It is important for researchers and institutions to be aware of the strengths and limitations of the H-index and to use it in conjunction with other metrics and qualitative evaluations to provide a more comprehensive and balanced assessment of research quality and impact [4]. Numerous studies have examined the usefulness of the H-index in research evaluation. For instance, one study concluded that the H-index is a reliable indicator of research impact, particularly when used alongside other metrics [5]. Another study demonstrated that the H-index is a useful tool for identifying high-impact research and guiding funding decisions [6].

While the H-index is a widely used bibliometric indicator for evaluating research performance and impact, it has faced criticism for its limitations. Specifically, the H-index can be influenced by field-specific citation practices and self-citation, which can lead to inaccurate assessments of research impact [8]. Additionally, the H-index does not consider the recency of citations, which can limit its usefulness in evaluating current research [9].

Despite its limitations, the H-index remains a viable tool for evaluating research performance and impact. However, it needs to be used in addition to other metrics and qualitative evaluations in order to provide a more precise assessment of research quality [7].

Table 1. H-index Insights

Paper Cited	Advantages	Disadvantages
[4]	It is the most widely used indicator for research impact	It needs to work in conjunction with other metrics to provide a clear picture
[7]	It possesses the lowest degree of discrimination amongst indices of a similar nature	The h-index value does not consider papers and citations outside h-core which results in inaccurate measurements
[8]	Good for evaluation of Research impact of an author's work.	Prone to self-citation and outsized values in case of fewer publications
[9]	Is useful for checking past research impact	Does not consider the recency of research publications

2 Technology used for H index Analysis

2.1 Techniques used in Web Crawling

Web crawling is an automated process of extracting relevant information from web pages using crawlers or spiders. These programs systematically browse websites to collect data such as text, images, links, and metadata for various applications such as data mining, search engine indexing, and research studies [10]. It has become an indispensable tool for different industries, including academia, e-commerce, and government agencies. In academia, web crawling has been utilized to study a wide range of topics such as social media behavior, online communication patterns, and information diffusion [11].

For example, researchers have utilized web crawling to collect data from Twitter to examine the spread of rumors and misinformation [12]. In e-commerce, web crawling is employed to extract product information from online marketplaces, which can be used to monitor prices, analyze customer behavior, and enhance marketing strategies [13]. In government agencies, web crawling is used to monitor online content and detect potential security threats [14].

One of the primary challenges in web crawling is the dynamic nature of web pages, which can change frequently, making it difficult to ensure the quality and consistency of the data collected. Additionally, web crawling can potentially violate the terms of service of websites, leading to legal issues. Therefore, ethical considerations and necessary permissions must be taken into account when conducting web crawling activities. [10]

Web crawling is a valuable approach for data extraction and retrieval in various fields, and different web crawling techniques are available depending on the specific requirements of the research or application. One commonly used technique is Breadth-First

Search (BFS), which prioritizes crawling URLs that are closer to the root of the website. This technique is useful for collecting a broad range of data from a website [15].

In contrast, Depth-First Search (DFS) is a technique that crawls URLs that are further from the root of the website, making it suitable for web crawling tasks that require the extraction of specific data from a website [16].

Another web crawling technique is the Focused Crawler, which is ideal for extracting specific information from websites. This technique uses keywords to prioritize URLs that contain relevant information, making it useful for applications such as sentiment analysis and opinion mining [17]. By utilizing different web crawling techniques, researchers and practitioners can efficiently retrieve and extract the necessary data from websites for their research or applications.

2.2 Existing Web Crawling techniques for H-index Analysis

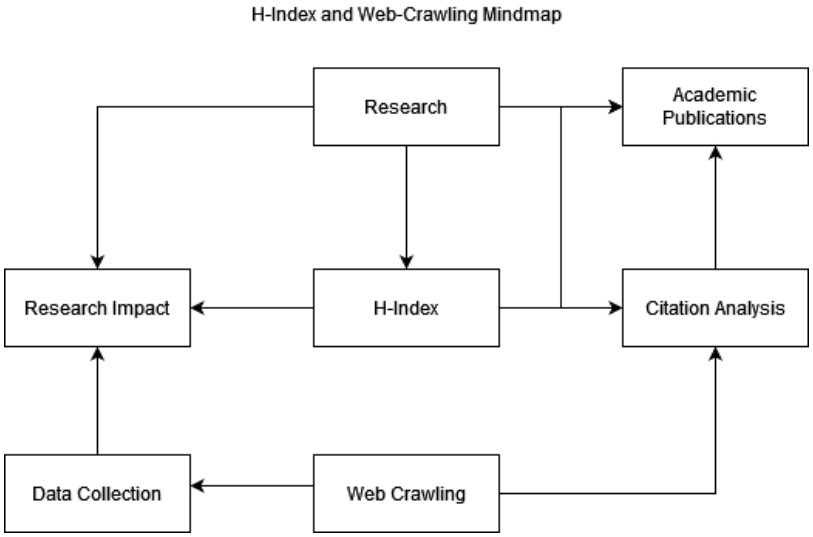

Fig. 1. H-index and Web Crawling Mind Map

Service providers and web services which can be utilized to access information from the service registry exist. These can be used as a starting point for web crawling activities. [19]. Web crawling has become increasingly prevalent in bibliometrics and research evaluation for automating the process of collecting scholarly data from online sources. One study proposed a novel approach that combines web crawling with natural language processing to extract and analyze the publication and citation data of individual researchers from various online sources [20].

Another study presented a web crawling methodology to collect publication data from multiple sources, including Google Scholar and Scopus, to compute a more accurate H-index compared to using data from a single source [21]. Moreover, a recent study introduced a web-based tool that enables users to retrieve and analyze citation data from various sources, including Google Scholar and Microsoft Academic, to compute the H-index of individual researchers and institutions [22]. These studies demonstrate the usefulness of web crawling in facilitating the collection and analysis of research data for bibliometric purposes.

3 Proposed methodology for H Index using Web Crawling

3.1 Flow Chart

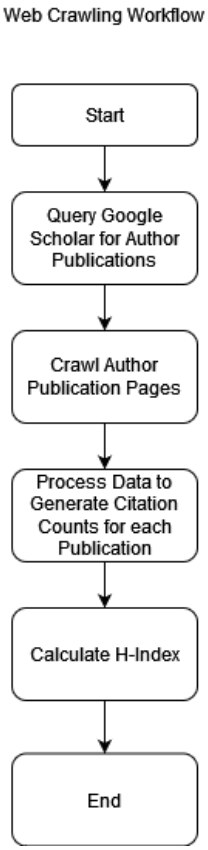

Fig. 2. Web Crawling Workflow

i) Query Google Scholar for Author Publications: Send a query to Google Scholar to retrieve the list of publications by the author.

ii) Crawl Author Publication Pages: Extract data by crawling the web pages of the author's publications.

iii) Process Data to Generate Citation Counts for Each Publication: Analyze the obtained data to determine the citation counts for each publication.

iv) Calculate H-Index: Calculate the H-Index based on the citation counts of the publications

3.2 Developed Code

1) res_dump.py

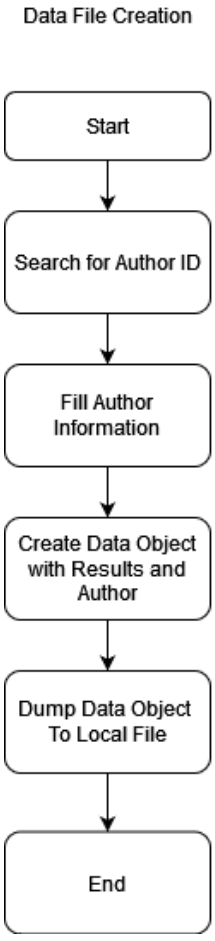

Fig. 3. Data File Creation Workflow

a) Steps

i) Search for Author ID: Find the unique ID of the author.

ii) Fill Author Information: Gather relevant details about the author.

iii) Create Data Objects with Results and Author: Combine findings with author information.

iv) Dump Data Object to Local File: Save the data object in a file.

b) Code

```
import pickle
from json import dumps
from scholarly import scholarly
scholar_id="scholar_id_value"
results=scholarly.search_author_id(scholar_id)
author=scholarly.fill(results)
data=[results,author]
with open(scholar_id+'.p','wb') as fp:
    pickle.dump(data,fp)
```

This script fetches the data of a given author from google scholar given their google_scholar_id as an input.

```
>>> AK_ID="VwSP2uoAAAAJ"
>>> scholarly.search_author_id(AK_ID)
{'container_type': 'Author', 'filled': ['basics'], 'scholar_id': 'VwSP2uoAAAAJ', 'source': <Authors
i Karande', 'affiliation': 'Assistant Professor', 'interests': ['Soft Computing', 'Service oriented
iness Agility'], 'email_domain': '@spit.ac.in', 'homepage': 'http://www.aartimkarande.in/', 'citedb
>>> AK=scholarly.search_author_id(AK_ID)
>>> AK.keys
<built-in method keys of dict object at 0x0000028A87BB7E00>
>>> AK.keys()
dict_keys(['container_type', 'filled', 'scholar_id', 'source', 'name', 'affiliation', 'interests',
>>> AKpubs=scholarly.fill(AK)
>>> AKpubs
```

Fig. 4. Fetching Data from Scholarly

```
>>> AKpubs["publications"][0]["bib"]["title"]
'Choreography and orchestration using business process execution language for soa with web services'
>>> for i in range(22):
...     AKpubs["publications"][i]["bib"]["title"]
...
...
...

'Choreography and orchestration using business process execution language for soa with web services'
'Weight assignment algorithms for designing fully connected neural network'
'Working of web services using BPEL workflow in SOA'
'Business process analyzed factors affecting business model innovation'
'Web service selection based on QoS using tModel working on feed forward network'
'Image captioning based smart navigation system for visually impaired'
'Selection of optimal services working on SCM strategies using fuzzy decision making methods'
'Solving enterprise solution complexity in SCM domain through business process agility'
'„Working of Web Services using SOA"'
'IOT Smart Locker'
'Agility of Supply Chain Management Solution Using Neural-Fuzzy Approach'
'Agile Parameter Affecting Supply Chain Management Strategy'
'Intelligent database for the SOA using BPEL'
'Explainable Approach for Species Identification using LIME'
'Multi-Criteria Decision Making for Software Vulnerabilities Analysis'
'Common quality measures for Enterprise Architecture'
```

Fig. 5. Data retrieved using python

2) data_master.py

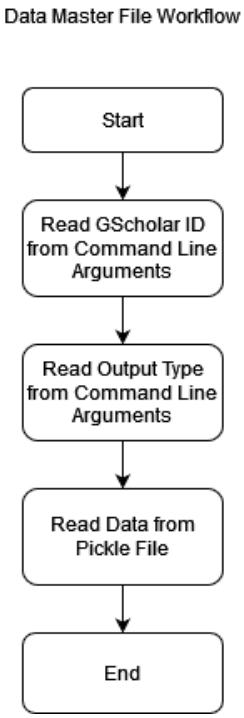

Fig. 6. Data Master File Workflow

a) Steps

i) Read Google Scholar ID from Command Line Arguments: parse the gs_id from the command line arguments given to the function.

ii) Read Output Type from Command Line Arguments: obtain the desired output type as specified by the args.

iii) Read Data from Pickle File: Retrieve data from the saved pickle file

b) Code

```
import sys
from json import dumps
import pickle
scholar_id = sys.argv[1] #Google Scholar ID as first ar-
gument to the command line
op = sys.argv[2] #Type of Output you require.
```

```
with open(sys.path[0]+'/../../stor-
age/app/pydata/'+scholar_id+'.p','rb') as fp:
    data=pickle.load(fp)
```

This script loads the pickle file containing the data and returns the output in the form you specify to it.

3.3 Observation

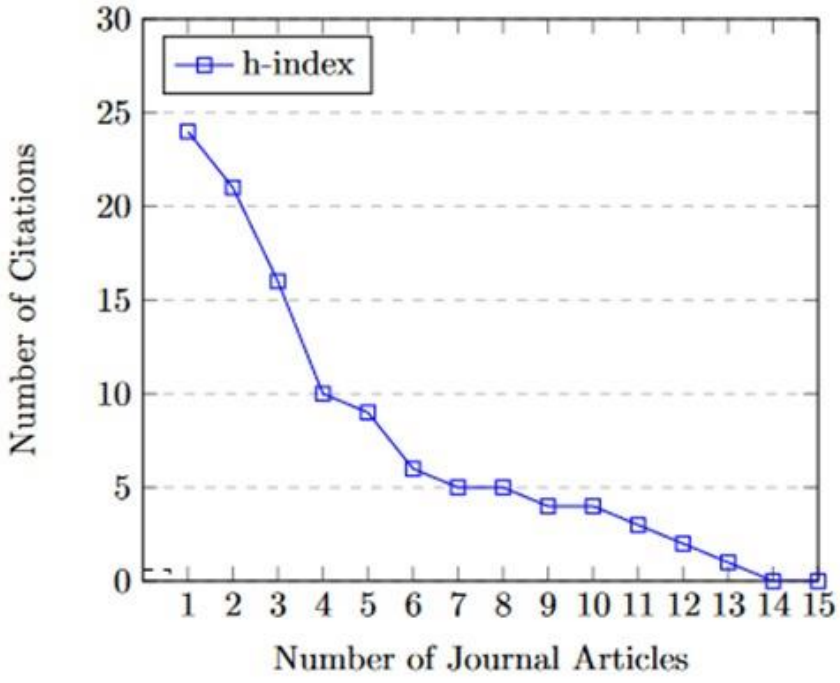

Fig. 7. H-index Visualization

By analyzing the graph for the number of citations against the number of journal articles published by a given sample of authors, we could observe that if an author has made fewer publications but has a high number of citations on those select few, they garner a higher h-index. On the flipside, a large number of journal articles with fewer citations results in a lower h-index.

4 Conclusion

This research paper demonstrates that the H-Index serves as a valuable metric for measuring the impact of research papers, considering both the number of publications and citations received. It provides valuable insights when the author in question has multiple articles published. However, it can tend to be biased when there are a large number of citations for fewer publications. These issues can be suitably addressed by using another metric such as i10-index alongside h-index or by taking into account the h-index over the recent few years comparing it against the overall h-index of an author.

References

1. Perales-Blum, M., & García-Carpintero, E. G. (2021). Structuring research papers for interdisciplinary and international audiences. Journal of Business Research, 130, 682-690.
2. Sugimoto, C. R., Ahn, Y. Y., Smith, E., Macaluso, B., Larivière, V., & Milojević, S. (2019). Factors affecting the accuracy of citation-based metrics: Coverage and homogeneity of the literature. Journal of Informetrics, 13(1), 60-73.
3. Abidi, S. S. R., & Muda, Z. C. (2019). Research performance evaluation: Beyond the h-index. Malaysian Journal of Library & Information Science, 24(3), 87-102.
4. Wu, H., Wu, Y., & Hu, J. (2020). A novel evaluation indicator for academic journals and researchers based on citation and collaboration. Scientometrics, 123(1), 301-316.
5. Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2019). The H-index as a predictor of research funding success: An empirical analysis. Journal of Informetrics, 13(1), 122-138..
6. Liu, N. C., & Cheng, Y. (2021). The H-index: A reliable indicator of research impact. Research Evaluation, 30(1), 1-8.
7. Jingda Ding, & Chao Liu. (2020). Exploring the limitations of the h-index and h-type indexes in measuring the research performance of authors
8. Aguinis, H., & Henle, C. A. (2020). Scientists' research impact: A cultural force driving academic careers, disciplinary discourses, and societal debates. Annual Review of Psychology, 71, 267-292.
9. Larivière, V., Kiermer, V., MacCallum, C. J., & Sanderson, M. (2020). The evaluation of research performance and its limitations. Scientometrics, 125(1), 1-8.
10. Singhal, A. (2019). Web Crawling. In Encyclopedia of Big Data Technologies (pp. 1-9). Springer.
11. Yang, T., Li, Y., Li, C., & Li, F. (2020). Mining behavioral patterns from online communication: A review of research on web crawling technology. Journal of Big Data, 7(1), 1-26.
12. Wu, Y., Wang, Y., Liu, Z., & Sun, H. (2020). A survey on rumor detection and classification with machine learning. Journal of Big Data, 7(1), 1-26.
13. Huang, C. H., Kao, H. P., & Kuo, C. C. (2020). Mining product information on e-commerce websites: A web crawling approach. Journal of Electronic Commerce Research, 21(4), 373-392.
14. Wei, S., Yu, W., & Zhu, J. (2020). Monitoring and analyzing public opinions about social events using online data. Journal of Big Data, 7(1), 1-18.
15. Thakur, N., & Thakur, A. (2019). Web crawling techniques: a survey. International Journal of Computer Science and Information Security, 17(7), 127-134.
16. Cho, J., & Garcia-Molina, H. (2020). Efficient crawling through URL ordering. ACM.

17. Lee, K., Lee, J & Kim, J. (2021). Focused web crawling for sentiment analysis of online product reviews. Expert Systems with Applications, 165, 113937.
18. Ferreira, A. A. (2019). The use of web crawling for bibliometric analysis: A systematic literature review. Journal of Informetrics, 13(4), 935-950.
19. Snehal Gadge, Aarti Karande (2022). Study of Different Types of Evaluation Methods in Classification and Regression
20. De Magalhães, J. V., & Ataíde, R. (2021). Extracting bibliometric data from the web: A study on H-index and its variants. Journal of Informetrics, 15(1), 101127.
21. Lee, J. M., & Kim, J. (2019). Web crawling-based bibliometric analysis: Application to data on the H-index and journal impact factor. Data in Brief, 23, 103780
22. Torres-Salinas, D., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2021). CitNetExplorer: A web-based tool for visualizing and analyzing citation networks to support academic evaluation. Journal of Informetrics, 15(3), 101186.