

**DESIGN AND DEVELOPMENT OF A CITATION  
ANALYSIS SYSTEM BASED ON DISTRIBUTED LEDGER  
TECHNOLOGY**

Thesis Submitted for the Award of the Degree of

**DOCTOR OF PHILOSOPHY**

**in**

**Computer Applications**

**By**

**Parul Khurana**

**Registration Number: 41500037**

**Supervised By**

**Dr. Gulshan Kumar (16865)**

**Computer Science and  
Engineering (Associate Professor)  
Lovely Professional University**

**Co-Supervised by**

**Dr. Geetha Ganesan**

**Professor and Dean –  
Faculty and Delivery,  
Jain University**

**Co-Supervised by**

**Dr. Kiran Sharma**

**Engineering & Technology  
(Assistant Professor)  
BML Munjal University**



**L** OVELY  
**P** ROFESSIONAL  
**U** NIVERSITY

*Transforming Education Transforming India*

**LOVELY PROFESSIONAL UNIVERSITY, PUNJAB**

**2023**

---

# DECLARATION

I declare that the thesis entitled "Design and development of a citation analysis system based on Distributed Ledger Technology" has been prepared by me under the guidance of Dr. Gulshan Kumar, Associate Professor, School of Computer Science and Engineering, Lovely Professional University, India, and under the co-guidance of Dr. Geetha Ganesan, Professor and Dean – Faculty and Delivery, Jain University, India, and Dr. Kiran Sharma, Assistant Professor, School of Engineering and Technology, BML Munjal University, India. No part of this thesis has formed the basis for the award of any degree or fellowship previously.

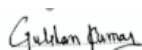
*Parul*

Parul Khurana  
School of Computer Applications  
Lovely Professional University  
Jalandhar – Delhi G.T.Road (NH-1)  
Phagwara, Punjab – 144411  
India  
Date: February 6, 2023

---

# CERTIFICATE

This is to certify that the thesis entitled "Design and development of a citation analysis system based on Distributed Ledger Technology", which is being submitted by Mr. Parul Khurana for the award of the degree of Doctor of Philosophy in Computer Applications from the Faculty of Engineering and Technology, Lovely Professional University, Punjab, India, is entirely based on the work carried out by him under our supervision and guidance. The work reported, embodies the original work of the candidate and has not been submitted to any other university or institution for the award of any degree or diploma, according to the best of our knowledge.



Dr. Gulshan Kumar

Associate Professor, School of Computer Science and Engineering,  
Lovely Professional University, Phagwara, Punjab-144411, India



Dr. Geetha Ganesan

Professor and Dean – Faculty and Delivery,  
Jain University, Bengaluru, Karnataka-560069, India



Dr. Kiran Sharma

Assistant Professor, School of Engineering and Technology,  
BML Munjal University, Gurugram, Haryana-122413, India

Date: February 6, 2023



*“The aim of life is inquiry into the Truth ...”*

– Bhagavata Purana

~ dedicated to my family ~

---

# ACKNOWLEDGEMENT

In the first place, I would like to thank my supervisor, Dr. Geetha Ganesan, for her constant guidance, support, and encouragement in the establishment, during, and completion of my Ph.D. research work. I also express my sincere and deepest gratitude to my second supervisor, Dr. Gulshan Kumar, for his constant suggestions and assistance in my Ph.D. research work. I also wish to express my immense gratitude to my third supervisor, Dr. Kiran Sharma, without whose guidance this Ph.D. research work was not possible.

My special thanks to Prof. Rajeev Sobti, Prof. Navdeep Dhaliwal, and Dr. Amandeep Bagga for their constant motivation in this new chapter of my life. I also express my immense gratitude to Mr. Enrico Bacis, Mr. Michael E. Rose, Dr. Ranbir Singh Batth, Mr. Sukhvir Singh, and Northwestern University, USA for providing the resources required for scrapping the data for my Ph.D. work.

I express my sincere and deepest gratitude to Lovely Professional University for giving me every opportunity to pursue my Ph.D.

Words cannot express how grateful I am to my mother, Mrs. Parmod Khurana, and my father, Mr. Kuldip Kumar Khurana, for all the sacrifices that they have made on my behalf. I am forever grateful to my wife, Mrs. Harmanjit Kaur, and my son, Master Khushaan Khurana, for always being there in my happy and difficult times. They are the constant source of inspiration in every endeavour of my life.

Last but not least, I dedicate this thesis to my family with love and gratitude.

Finally, I thank the Lord Krishna for all I have been given and gifted with.

Date: February 6, 2023

*Parul*  
Parul Khurana

---

# Abbreviations

---

Abbreviations	Description
<b>AHCI</b>	Arts & humanities citation index
<b>AICTE</b>	All India council for technical education
<b>API</b>	Application programming interface
<b>BKCI</b>	Book citation index
<b>CPCI</b>	Conference proceedings citation index
<b>CSV</b>	Comma separated values
<b>DLT</b>	Distributed ledger technology
<b>DOI</b>	Digital object identifier
<b>ESCI</b>	Emerging sources citation index
<b>EVM</b>	Ethereum virtual machine
<b>ID</b>	Identifier
<b>IEEE</b>	Institute of electrical and electronics engineers
<b>IEST</b>	Indian institute of engineering science and technology
<b>IIT</b>	Indian institute of information technology
<b>IIM</b>	Indian institute of management
<b>IISC</b>	Indian institute of science
<b>IISER</b>	Indian institute of science education and research
<b>IIT</b>	Indian institute of technology
<b>IoT</b>	Internet of things
<b>IP</b>	Internet protocol
<b>ISBN</b>	International standard book number

<b>ISSI</b>	International society for scientometrics and informetrics
<b>ISSN</b>	International standard serial number
<b>MHRD</b>	Ministry of human resource and development
<b>NIRF</b>	National institutional ranking framework
<b>NIT</b>	National institute of technology
<b>ORCID</b>	Open researcher and contributor Identifier
<b>ORG</b>	Organization
<b>PBI</b>	Proof of bibliometric indicators
<b>SCI</b>	Science citation index
<b>SSCI</b>	Social sciences citation index
<b>THE</b>	Times higher education
<b>UGC</b>	University grants commission
<b>URL</b>	Uniform resource locator
<b>WoS</b>	Web of science
<b>XML</b>	Extensible markup language

---

# ABSTRACT

Bibliometric studies reveal the usage of indexing databases such as Scopus, Web of Science, PubMed, Google Scholar, etc., to track the scientific progress of the research community. However, the choice of journals and indexing approach varies among these indexing databases, hence they produce different informetrics such as publications, citations, and  $h$ -index count for the same individual or groups. This creates a dilemma in the minds of stakeholders such as hiring agencies, accreditation agencies, funding agencies, and government bodies to select or reject provided informetrics. At the end, they are left with multiple informetrics for the same individual or group of individuals. At present, there is no common platform or system that can present or generate unified (single) informetrics for an individual or group across multiple indexing databases.

With the objective of generating unified (single) informetrics, a literature review was done to study the depth of the problem. It was found that various authors have compared indexing databases based on different parameters, such as their statistics and orientation, on the availability of digital object identifiers, coverage, strength, searching capabilities,  $h$ -index and their content comprehensiveness, etc., but none of the literature provides any comprehensive solution. Hence, a research gap was identified and a problem definition was prepared and finalized.

For valid ranking assessments,  $h$ -index is one of the most essential, robust, primitive, quantitative, and singular measures used to assess the quality, impact, influence, and relevance of an individual's or group's work. Hence, the use of different indexing databases to generate informetrics of an individual or a group also shows a significant impact on their ranking. To analyze the situation, data was extracted for three entities such as author, organization, and journal from Scopus and Web of Science.



In the next step, a process to generate resultant database named as “Conflate”, was initiated. DOI based filtration for publications and citations was applied and citation analysis with weight assignment was performed based on informetrics such as publications, citations and  $h$ -index of author, organization and journal. We did a useful investigation based on the  $h$ -index information extracted from multiple indexing databases, and identified the possible improvements we can make in the form of  $h_c$  as a simple compliment to  $h$ -index.

After generating the required informetrics, the idea of using distributed ledger technology was introduced with the mapping of distributed ledger technology with informetrics. Available literature on distributed ledger technology was studied to identify the existence of this technology in various domains. The concept of consensus, a robust feature of distributed ledger technology, was also elucidated with the introduction of new consensus mechanisms such as proof of bibliometric indicators. In the end, the resultant database, named “Conflate” was deployed to fit into the insight of bibliometrics with the use of an Ethereum-oriented distributed ledger-based system with Truffle and Ganache.

Hence, in the current scenario in 2022, where an individual or a group provides multiple informetrics to its stakeholders, this novel approach has overcome this limitation and generates a single informetrics in the form of publications, citations, and  $h$ -index for an individual or a group. Moreover, the use of distributed ledger technology in the implemented system has given greater transparency and instant automation to the generated informetrics as well. The key findings of the present work may be summarized as follows: (i) It offers a unified approach to keep the track of informetrics associated with author, organization, and journal. (ii) Mapping of numerous indexing databases for the calculation of the absolute number of publications, citations and  $h$ -index. (iii) The implemented system makes it easier for its stakeholders to utilize a framework for a transparent, accurate, and simulated environment for measuring entities for different studies. (iv) Detailed assessments of key elements such as publications, citations, and  $h$ -index are provided. (v)  $h_c$  is introduced as a complementary approach to the  $h$ -index. (vi) Presentation of unified informetrics with robust, validated records and with the power of distributed ledger technology.

---

# CONTENTS

DECLARATION	i
CERTIFICATE	ii
ACKNOWLEDGEMENTS	iii
ABBREVIATIONS	iv
ABSTRACT	vi
CONTENTS	ix
LIST OF FIGURES	xiii
LIST OF TABLES	xvii
<b>1 Introduction</b>	<b>1</b>
1.1 Indexing databases and their context . . . . .	2
1.2 Usefulness of citations . . . . .	4
1.3 Informetrics in science of science . . . . .	6
1.4 Motivation behind the work . . . . .	7
1.5 Contribution based on IEEE taxonomy 2021 . . . . .	8
1.6 Outline of the thesis . . . . .	9
<b>2 Review of literature</b>	<b>13</b>
2.1 Indexing databases . . . . .	13
2.2 Studies related to comparative analysis of indexing databases . . . . .	15
2.3 Revolution of distributed ledger technology . . . . .	18
2.4 Distributed ledger technology based applications . . . . .	19
2.5 Consensus mechanism in distributed ledger technology . . . . .	20
2.6 Research gap . . . . .	20

---

2.7	Discussion and summary . . . . .	28
<b>3</b>	<b>Indexing databases and informetrics</b>	<b>29</b>
3.1	Impact of informetrics on authors . . . . .	29
3.2	Author's ranking in Scopus and WoS based on informetrics . . . . .	33
3.3	Comparative analysis of informetrics . . . . .	34
3.4	Impact of indexing databases on author's rankings . . . . .	36
3.5	Discussion and summary . . . . .	37
<b>4</b>	<b>Linking of indexing databases and generation of unified informetrics (UI)</b>	<b>39</b>
4.1	Entity specification and linkage of citation analysis . . . . .	40
4.1.1	Author level bibliometrics . . . . .	41
4.1.2	Organization level bibliometrics . . . . .	42
4.1.3	Journal level bibliometrics . . . . .	43
4.1.4	Entity identifiers . . . . .	43
4.1.5	Digital object identifiers . . . . .	44
4.1.5.1	Author level bibliometrics . . . . .	46
4.1.5.2	Organization level bibliometrics . . . . .	47
4.1.5.3	Journal level bibliometrics . . . . .	49
4.2	Methodology . . . . .	51
4.2.1	Generation of doi based citation database . . . . .	52
4.2.2	Computation of weighted unified informetrics . . . . .	53
4.2.3	The weighted unified informetrics algorithm . . . . .	54
4.3	Data description and filtering . . . . .	54
4.3.1	Data sources . . . . .	54
4.3.2	Data analytic . . . . .	58
4.3.3	Article extraction and filtration . . . . .	58
4.3.3.1	Author level bibliometrics . . . . .	59
4.3.3.2	Organization level bibliometrics . . . . .	60
4.3.3.3	Journal level bibliometrics . . . . .	61
4.3.4	Citation extraction and filtration . . . . .	61
4.3.4.1	Author level bibliometrics . . . . .	62
4.3.4.2	Organization level bibliometrics . . . . .	63
4.3.4.3	Journal level bibliometrics . . . . .	64
4.4	Citation analysis and unified informetrics . . . . .	65
4.4.1	Author level bibliometrics . . . . .	65
4.4.2	Organization level bibliometrics . . . . .	66
4.4.3	Journal level bibliometrics . . . . .	67
4.5	Discussion and summary . . . . .	68
<b>5</b>	<b>Statistical analysis of Conflate (unified informetrics (UI))</b>	<b>70</b>

---

5.1	Author level bibliometrics . . . . .	70
5.1.1	Number of publications . . . . .	73
5.1.2	Number of citations . . . . .	74
5.1.3	Measuring the $h$ -index . . . . .	76
5.1.4	Self-citations vs. total-citations . . . . .	77
5.1.5	Repeated-citations vs. total-citations . . . . .	78
5.1.6	Actual-citations vs. total-citations . . . . .	79
5.1.7	No. of citations vs. average $h$ -index . . . . .	80
5.2	Organization level bibliometrics . . . . .	81
5.2.1	Number of publications . . . . .	83
5.2.2	Number of citations . . . . .	83
5.2.3	Measuring the $h$ -index . . . . .	86
5.2.4	Self-citations vs. total-citations . . . . .	88
5.2.5	Repeated-citations vs. total-citations . . . . .	89
5.2.6	Actual-citations vs. total-citations . . . . .	90
5.2.7	No. of citations vs. average $h$ -index . . . . .	91
5.3	Journal level bibliometrics . . . . .	91
5.3.1	Number of publications . . . . .	93
5.3.2	Number of citations . . . . .	95
5.3.3	Measuring the $h$ -index . . . . .	96
5.3.4	Self-citations vs. total-citations . . . . .	97
5.3.5	Repeated-citations vs. total-citations . . . . .	99
5.3.6	Actual-citations vs. total-citations . . . . .	100
5.3.7	No. of citations vs. average $h$ -index . . . . .	100
5.3.8	Measuring the impact factor . . . . .	101
5.4	Discussion and summary . . . . .	104
<b>6</b>	<b>Distributed ledger technology based implementation of Conflate</b>	<b>106</b>
6.1	Analogy and consensus for applying DLT . . . . .	107
6.1.1	Mapping of distributed ledger technology with research publi- cations . . . . .	108
6.1.2	Proof of bibliometric indicators (PBI) - consensus mechanism .	109
6.1.2.1	Design of DLT based system for unified informetrics .	109
6.1.2.2	Design of PBI consensus mechanism . . . . .	110
6.2	Implementation details . . . . .	111
6.2.1	Input and output demonstration of informetrics in blockchain .	111
6.2.2	Entity registration using Identifiers . . . . .	112
6.2.2.1	Author level bibliometrics . . . . .	112
6.2.2.2	Organization level bibliometrics . . . . .	113
6.2.2.3	Journal level bibliometrics . . . . .	113

---

6.2.3	Block and transaction generation and confirmation process for all entities . . . . .	114
6.2.4	Fetching informetrics stored in blockchain . . . . .	115
6.2.4.1	Author level bibliometrics . . . . .	115
6.2.4.2	Organization level bibliometrics . . . . .	116
6.2.4.3	Journal level bibliometrics . . . . .	118
6.3	Discussion and summary . . . . .	118
<b>7</b>	<b>Conclusion and outlook</b>	<b>120</b>
7.1	Concluding remarks . . . . .	120
7.2	How our work is overcoming the identified research gap during literature review? . . . . .	124
7.3	Summary of contribution . . . . .	125
7.4	Future direction . . . . .	125
	<b>REFERENCES</b>	<b>128</b>
	<b>PUBLICATIONS AND PATENTS</b>	<b>145</b>

---

# LIST OF FIGURES

1.1	Contribution based on IEEE taxonomy 2021. . . . .	9
3.1	Author ranking using the $h$ , $g$ , and $h_c$ indices. . . . .	35
3.2	Correlation coefficients of (a) $h$ and $h_c$ , and (b) $h$ and $g$ for both Scopus and WoS. . . . .	36
3.3	For $h = 2$ , an example of the $h_c$ growth curve based on $H_{cite}$ . . . . .	36
3.4	For (a) Scopus and (b) WoS, the fraction of authors, $h$ and $h_c$ wise. . .	37
4.1	Comparative analysis of 400 authors based on DOI information. . . . .	47
4.2	Comparative analysis of 100 organizations based on DOI information. .	49
4.3	Comparative analysis of 1000 journals based on DOI information. . . .	51
4.4	Flowchart demonstrates the computation of weighted unified informetrics	55
4.5	Schematic representation of the proposed weighted unified informetrics	56
4.6	Flowchart demonstrates the process of visiting the author's, organization's, and journal's profile. . . . .	57
4.7	Representation of concepts used to retrieve, compile, analyze and present the unified informetric ledger - Conflate. . . . .	58
5.1	Filtration process listing all the steps, from random author profiles to final list of author profiles. . . . .	72
5.2	Discipline/Subject area details of 400 authors. . . . .	72
5.3	A comparison of publications of 400 authors based on Scopus, Web of Science and Conflate. . . . .	73
5.4	Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level. . . . .	74

---

5.5	A comparison of citations of 400 authors based on Scopus, Web of Science and Conflate. . . . .	75
5.6	Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level. . . . .	75
5.7	A comparison of $h$ -index of 400 authors based on Scopus, Web of Science and Conflate. . . . .	76
5.8	Comparative analysis based on $h$ -index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level. . . . .	77
5.9	A comparison of total and self citations of 400 authors based on Scopus, Web of Science and Conflate. . . . .	78
5.10	A comparison of total and repeated citations of 400 authors based on Scopus, Web of Science and Conflate. . . . .	79
5.11	A comparison of total, self, repeated and actual citations of 400 authors based on Scopus, Web of Science and Conflate. . . . .	80
5.12	A comparison of citations and $h$ -index of 400 authors based on Scopus, Web of Science and Conflate. . . . .	81
5.13	Filtration process listing all the steps, from initial selection of platform to final list of organization profiles. . . . .	82
5.14	Details of 100 organizations on the basis of their type. . . . .	83
5.15	A comparison of publications of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	84
5.16	Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level. . . . .	84
5.17	A comparison of citations of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	85
5.18	Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level. . . . .	86
5.19	A comparison of $h$ -index of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	87
5.20	Comparative analysis based on $h$ -index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level. . . . .	87

---

5.21	A comparison of total and self citations of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	88
5.22	A comparison of total and repeated citations of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	89
5.23	A comparison of total, self, repeated and actual citations of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	90
5.24	A comparison of citations and $h$ -index of 100 organizations based on Scopus, Web of Science and Conflate. . . . .	91
5.25	Filtration process listing all the steps, from initial selection of journals to final list of journals. . . . .	93
5.26	Details of 1000 journals on the basis of their disciplines or subject areas.	93
5.27	A comparison of publications of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	94
5.28	Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level. . . . .	94
5.29	A comparison of citations of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	95
5.30	Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level. . . . .	96
5.31	A comparison of $h$ -index of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	97
5.32	Comparative analysis based on $h$ - index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level. . . . .	98
5.33	A comparison of total and self citations of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	98
5.34	A comparison of total and repeated citations of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	99
5.35	A comparison of total, self, repeated and actual citations of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	100
5.36	A comparison of citations and $h$ -index of 1000 journals based on Scopus, Web of Science and Conflate. . . . .	101
5.37	Filtration process listing all the steps, from initial selection of journals to final list of journals for impact factor calculation. . . . .	102



---

5.38	Details of 746 journals on the basis of their disciplines or subject areas.	103
5.39	A comparison of impact factor of 746 journals based on sources (Scopus and Web of Science together) and Conflate. . . . .	103
5.40	Comparison of impact factor on the basis of results generated by Conflate and sources from Scopus and Web of Science. . . . .	104
6.1	Schematic representation of DLT mapping with research publications. .	109
6.2	Representation of system model of PBI. . . . .	110
6.3	Author registration process for (a) user interface (b) transaction (c) registration confirmation. . . . .	112
6.4	Organization registration process for (a) user interface (b) transaction (c) registration confirmation. . . . .	113
6.5	Journal registration process for (a) user interface (b) transaction (c) registration confirmation. . . . .	114
6.6	Block generated for an entity registration . . . . .	114
6.7	Transaction generated for a entity registration . . . . .	115
6.8	Number of confirmations received for a transaction generated for an entity registration . . . . .	116
6.9	Author retrieval for (a) particular author details (b) registered author's details. . . . .	117
6.10	Organization retrieval for (a) particular organization details (b) registered organization's details. . . . .	117
6.11	Journal retrieval for (a) particular journal details (b) registered journal's details. . . . .	118

---

# LIST OF TABLES

2.1	List of studies included comparative analysis, based on Scopus, Web of Science, Google Scholar etc. (sorted year wise) . . . . .	22
3.1	Demonstration of $h_c$ . . . . .	33
3.2	Statistics of $h$ , $h_c$ and $g$ . . . . .	34
3.3	Proportion of authors with varying $h$ and $h_c$ for Scopus (S) and WoS (W). . . . .	38
4.1	Comparative analysis of 400 authors based on DOI information. . . . .	47
4.2	Comparative analysis of 100 organizations based on DOI information. . . . .	50
4.3	Comparative analysis of 1000 journals based on DOI information. . . . .	52
5.1	Comparative analysis of publications - author level . . . . .	74
5.2	Comparative analysis of citations - author level . . . . .	76
5.3	Comparative analysis of average $h$ -index - author level . . . . .	77
5.4	Comparative analysis of self citations vs. total citations - author level . . . . .	78
5.5	Comparative analysis of repeated citations vs. total citations - author level . . . . .	79
5.6	Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - author level . . . . .	80
5.7	Comparative analysis of no. of citations vs. average $h$ -index - author level . . . . .	81
5.8	Comparative analysis of publications - organization level . . . . .	85
5.9	Comparative analysis of citations - organization level . . . . .	86
5.10	Comparative analysis of average $h$ -index - organization level . . . . .	88

---

5.11 Comparative analysis of self citations vs. total citations - organization level . . . . .	89
5.12 Comparative analysis of repeated citations vs. total citations - organization level . . . . .	90
5.13 Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - organization level . . . . .	90
5.14 Comparative analysis of no. of citations vs. average $h$ -index - organization level . . . . .	92
5.15 Comparative analysis of publications - journal level . . . . .	95
5.16 Comparative analysis of citations - journal level . . . . .	96
5.17 Comparative analysis of average $h$ -index - journal level . . . . .	98
5.18 Comparative analysis of self citations vs. total citations - journal level .	99
5.19 Comparative analysis of repeated citations vs. total citations - journal level . . . . .	100
5.20 Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - journal level . . . . .	101
5.21 Comparative analysis of no. of citations vs. average $h$ -index - journal level . . . . .	102

---

---

# CHAPTER 1

---

## Introduction

Research is a systematic investigation that gathers and analyzes the information to produce or contribute to generalized knowledge. It aims to expand human understanding of the physical, biological, or social environment beyond what is currently known. Because it employs a methodical approach known as the scientific method, research differs from other types of information discovery, and it represents an unseen opportunity for the betterment of humanity.

The comprehensive sharing of research work through scientific publishing in the academic society is possible globally with the existence of indexing agencies such as Elsevier's Scopus, Clarivate's Web of Science, PubMed, Google Scholar, and Microsoft Academic, etc. Enhanced visibility and fostering collaboration are also empowered by these indexing agencies. Due to the availability of such indexing agencies, research workflow among various stakeholders has emerged as a mixture of comprehensive scientific data and analytical tools. Research quality and scientific output are the most important criterion among various stakeholders, such as ranking agencies, hiring agencies, accreditation agencies, government bodies, and funding organizations.

Hence, the importance of indexing databases has greatly increased as a main source of publication metadata and citation metrics.

Indexing databases support the influence, recognition, and contribution of the scientific work of an author, an organization, and a journal. Finding the relevant citation information among these indexing databases is an enormous task, although many tools are provided for this purpose. The present thesis is entitled as “Design and development of a citation analysis system based on distributed ledger technology”. In my thesis, the novel approach based on distributed ledger technology to deal with citations is introduced. This thesis describes the research work carried out in the last five years and the personal beliefs about the rapidly growing field of distributed ledger technology.

## 1.1 Indexing databases and their context

Scientific work of an author, organization, and a journal is indexed in various indexing databases. These indexing databases are the organized collections of various scientific works, like articles, books, conference publications, patent records, etc., of different entities. These indexing databases also provide various tools to visualize, analyze, and present the data in an easy, convenient and graphical manner to their stakeholders, whereas some of the indexing databases are open source and some are paid.

These days, there are number of indexing databases available for the indexing of scientific works. For example,

- Scopus (<https://www.scopus.com/home.uri>).
- Web of Science (<https://www.webofknowledge.com/>).
- Google Scholar (<https://scholar.google.com/>).
- PubMed (<https://pubmed.ncbi.nlm.nih.gov/>).
- Openaire (<https://www.openaire.eu/>).

- Mendely (<https://www.mendeley.com/>).
- Zenodo (<https://zenodo.org/>).

The Institute for Scientific Information created the first scientific citation indexes (ISI). The Science Citation Index (SCI) was first published in 1964, followed by the Social Sciences Citation Index (1973) and the Arts & Humanities Citation Index (1978). These citation indexes were made available online in 1997 under the term “Web of Science”. These citation indexes, as well as several new ones, including the conference proceedings citation index, book citation index, and emerging sources citation index, were renamed as the “Web of Science Core Collection” (from now on, WoS). The availability of this data was critical to the growth of quantitative science studies as a discipline [1].

Two new academic bibliographic data sources with citation data were launched in November 2004. The first one was Elsevier’s Scopus, which is a subscription-based database that indexes documents selectively (documents from a pre-selected list of publications). The second one was the search engine Google Scholar, which was introduced a few weeks after Scopus. Unlike WoS and Scopus, Google Scholar takes a more broad and automated approach, indexing any supposedly academic document that its crawlers can find and access on the web, including those behind paywalls due to publisher agreements. Furthermore, Google Scholar is a free service that provides users with access to a broad and diverse citation index.

Due to their research impact and good citation value, some of the indexing databases are very popular among stakeholders. Two very common names in the list are Scopus and Web of Science. These two indexing databases are always compared with other indexing databases for any kind of analysis on the citations, content comprehension and bibliometric values. To access the content on indexing databases like Scopus and Web of Science, a subscription is required. Such subscriptions are generally organization-based. So, to get access to subscribed indexing databases, a login from an organization is required. Access is often IP-based where an individual may have to enter his/her credentials like username and password to login and download

the required content in the form of articles, book series, and conference proceedings. Account creation on both Scopus and Web of Science may be done to save the results of accessed records on a permanent basis as well.

Indexing databases like Google scholar, Openaire, Mendely, and Zenodo are providing free access to the limited records. One has the option of signing up on the websites of these indexing databases so that an author profile may be created to browse and save the searched results in the profile itself. Some of the indexing databases are promoting open access to the scientific work so that more and more stakeholders can approach the scientific work of other authors to take advantage of the collaborations [2].

There are multiple ways to access the data on these indexing databases. One easiest way is to search the required author, article, organization or journal information directly on the website. An alternative way is to access the required data on these indexing databases with the help of Application Programming Interfaces (APIs). APIs help to retrieve large amounts of data in an easy manner. One can customize queries and retrieve the results accordingly. These results can be saved in the format of XML (Extensible Markup Language) and CSV (Comma-separated values) files on a permanent basis as an offline database in the system. Such offline databases may be used as per the requirement for any kind of visualization or analysis later on [3].

## 1.2 Usefulness of citations

Citations provide the mechanism to give recognition and acknowledgement to the scientific work of other authors. When a scientific work is cited in any publication or study, stakeholders who are reading the work, recognize the efforts in different ways, like what is the source of the information; who are the other scientists working on similar problems or fields; and what are the contributions of other scientists as well; and to provide an overall view and strength to an idea of scientific work done by the author. It also helps to distinguish the scientific work from other scientists [4].

Cited record may include the important fields of information. For example, for citation type "article", important fields may include the information of an author, title, journal, volume, number, pages, year, DOI, and keywords, etc. For citation type "book", important fields may include information about the title, the author, ISBN, series, year, publisher, and keywords, etc. For citation type "online", important fields may include the information of an author, title, url, addendum, and keywords, etc. [5]. Different citation types may include articles, chapters, conference proceedings, books, Ph.D. theses, master's theses, technical reports, online materials, unpublished work, etc.

For different disciplines, there are different ways to cite scientific work. So, according to the discipline of an author, a required citation style may be used to cite the work of other scientists [6]. Different authors, organizations and journals show the quality and depth of their scientific work in the form of citations. Any publication in the form of an article, book, or website that gets any recognition gets it in the form of citation. Any bibliometrics, scientometrics, and informetrics used to calculate the scientific impact of an author, organization, or journal also use the concept of citations. A very common bibliometric indicators like  $h$ -index, and impact factor also uses the concept of citations for the final calculation of indicator values.

Citations are always embedded in the scientific work of an author, organization, or journal. A section named "bibliography" always contains the complete description of citations. Wherever it is required in the main scientific work, such citations may be called and cited in different styles depending on the citations' reference. One can cite in the format of numbers or in the format of alphabets or in the format of alphanumeric characters as per the discipline and requirements of styles. It is often recommended to follow the same style throughout the main document so that a common sequence, syntax, and similarity may be maintained in the complete document.



### 1.3 Informetrics in science of science

Informetrics encompass bibliometrics and scientometrics. It is widely used as a metric term for information process, phenomena and retrieval theory [7]. It covers empirical studies, theoretical studies, characteristics and quantitative aspects of literature documents and potentially informative text from any scientific as well as social community [8]. Authors visualize informetrics as an extension of traditional bibliometrics and have introduced and compared informetrics with other metrics as well. Among all, informetrics is the most general of the three terms. However, informetrics, bibliometrics, and scientometrics are often used interchangeably by various authors.

The study of the quantitative elements of information in any form, not only records or bibliographies, and in any social group, not just scientists, is described as informetrics. It considers both informal and recorded communication as well as information needs. With the increased availability of documentary materials and the discourse electronic formats, such as machine-readable databases and, more lately, the internet, informetric research based on electronic data sets has become more frequent. As a result, the term "informetrics" is used as a broad term to encompass and employ various studies of information measurements that fall outside the scope of both bibliometrics and scientometrics [9].

It is a prevalent misconception that scientometrics is nothing more than the measurement of scientific performance based on publications and citations, or the compilation of cleaned-up bibliographies on study topics enhanced by citation data. Scientometrics is the study of the quantitative aspects of science as a discipline or as a source of income. It is a branch of science sociology with implications for science policy. It entails quantitative analyses of scientific activity, such as publication, and hence overlaps with bibliometrics to some extent [10].

Bibliometrics is defined as the quantitative examination of publications with the goal of determining certain types of phenomena. It includes the measuring of document attributes as well as document-related operations. It analyses and measures the output of scientific articles using mathematical and statistical approaches. Scientific

and technological subjects have accounted for the vast majority of bibliometric investigations. It's worth noting that the measuring of published scholarship and scientific research has had its own momentum and terminology. This sort of publication has become essential for library and information science as well as scholarly communication, ranging from statistical bibliography to bibliometrics to scientometrics and informetrics [11].

Researchers can evaluate literature to determine disciplinary traits, scholarly obsolescence, institutional linkages and relationships, and the types of materials that make up scholarly pursuits. In several branches of study, bibliometrics is utilized as a methodology, most notably to map the publication pattern in various disciplines. For example, bibliometrics is a must-have tool for historians researching a discipline's intellectual history and evolution [12].

These days, educational initiatives in the field of informetrics are very popular and actively presented in foreign countries. The International Society for Scientometrics and Informetrics (ISSI) has been regularly conducting conferences since 1987 to study quantitative approaches in informetrics, bibliometrics, and scientometrics (<https://www.issi-society.org/home/>). Journal of Scientometrics (<https://www.springer.com/journal/11192/>) and Journal of Informetrics (<https://www.sciencedirect.com/journal/journal-of-informetrics>) also reflect the growth and expansion of “informetrics” [13].

## 1.4 Motivation behind the work

For ranking universities based on research parameters, various agencies like NIRF (<https://www.nirfindia.org/Home>) and Times Higher Education (THE) (<https://www.timeshighereducation.com/>) use databases of their choice for research evaluation based on reputation surveys, research income, quality of publications, research influence and productivity etc. THE partnered with Elsevier in 2014 to have a deeper amount of the research data required for global rankings [14], THE was using Thomson Reuters. On the other hand, in India, NIRF gives equal weight to

indexing databases such as Scopus, and Web of Science etc. As all indexing databases carry their own special features and fields of information, it's not feasible to decide which indexing database is better than others. Hence, it is not appropriate to give more weightage to any specific database over others. However, a unified measure can be used on such databases, where one can see a single informetrics across different indexing databases for an author, organization and journal. Generated single informetrics will act as a single measure of evaluation for various agencies such as ranking agencies, accreditation bodies, hiring agencies, promotion agencies, and funding bodies to evaluate the research performance of an individual as well as a group. This gives us the motivation to work towards the generation of unified informetrics across multiple indexing databases.

## 1.5 Contribution based on IEEE taxonomy 2021

As per IEEE taxonomy, the scientific work done during Ph.D. contributes to the intersection of the following areas (see Figure 1.1):

- Computers and information processing -> Blockchain
- Computers and information processing -> Publishing -> Bibliometrics
- Computers and information processing -> Publishing -> Scientific publishing
- Computers and information processing -> Data integration
- Computers and information processing -> Data preprocessing
- Computers and information processing -> Data handling
- Computers and information processing -> Distributed information systems
- Mathematics -> Algorithms -> Algorithm design and theory -> Consensus algorithm
- Professional communication -> Databases -> Distributed databases

- Professional communication -> Databases -> Information analysis -> Decision analysis
- Professional communication -> Databases -> Information integrity
- Professional communication -> Databases -> Information resources
- Professional communication -> Databases -> Information retrieval

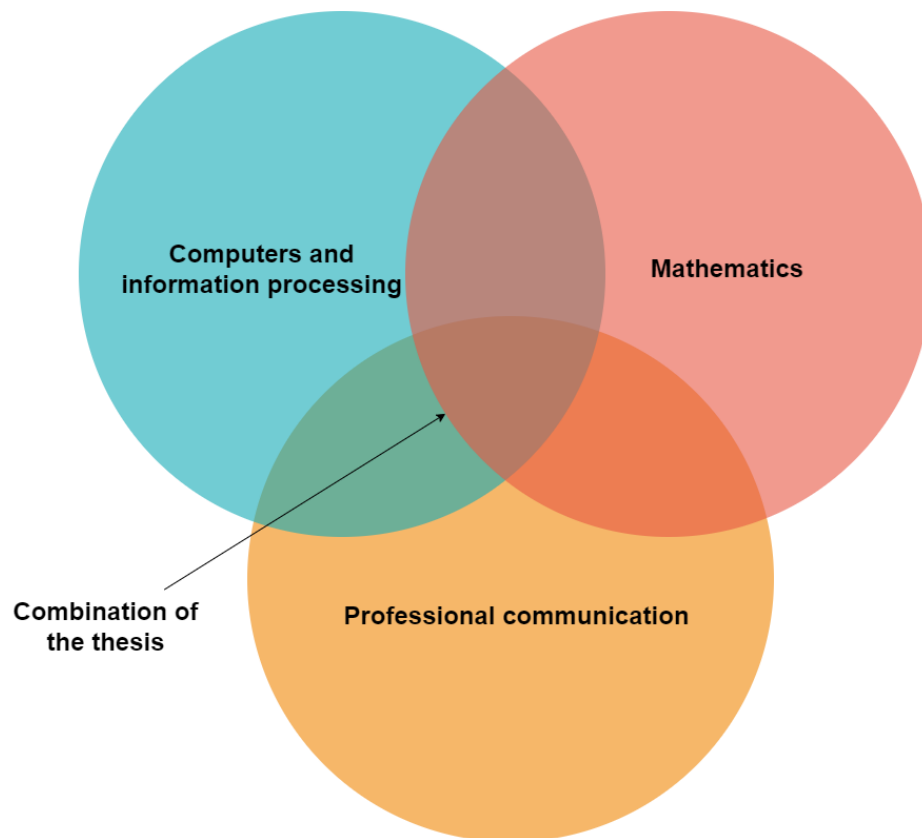


FIGURE 1.1: *Contribution based on IEEE taxonomy 2021.*

## 1.6 Outline of the thesis

The present thesis contains four research chapters, and the last chapter is the thesis conclusion and future directions. For every chapter, we have written an introduction, outcomes, and possible discussion. We can say that this thesis work is the

output of the joint guidance of the supervisor, co-supervisor, and the other collaborators. We have also tried to acknowledge the efforts of each and every individual or resource in the chapter itself.

**Chapter 2:** This chapter deals with the review of literature required to start, perform, and complete scientific work. At the beginning of this chapter, we talked about the literature available on indexing databases, based on their introduction and their comparative studies. In the next section of this chapter, we will talk about the literature available on distributed ledger technology, based on its introduction, its applications, and consensus mechanisms. In the last section, we introduced the research gap and the identified problem. The primary questions, answered in this chapter are:

- What is the current research on indexing databases?
- Why authors compared indexing databases based on various parameters?
- What are the limitations of the available literature on comparative analysis of indexing databases?
- What is the research gap and future scope based on available literature on comparative analysis of indexing databases?

**Chapter 3:** This chapter deals with the introduction of indexing databases, and  $h$ -index as an important informetrics in the scientific community with its features and limitations. At the beginning, we have discussed the concept of measuring informetrics based on different indexing databases. At the end, we have provided a supplementary approach named as  $h_c$  to overcome the limitations of  $h$ -index. The primary questions, answered in this chapter are:

- What is the impact of measuring informetrics with various indexing databases?
- How can we improve  $h$ -index by considering the weight of highest cited papers?

**Chapter 4:** This chapter deals with the core framework of the scientific work done to solve the identified problem. At the beginning of this chapter, we talked about

different entities we have taken to start with the solution of the problem. What is the relevance of taking such entities? How are those entities actually connected with scientific work? What is the importance of scientific work for those entities?.

In the next section of this chapter, we have explained the concept of different qualifiers that we have used to uniquely recognize these entities among different indexing databases. Furthermore, we have talked about the sources of data that were used for citation analysis. In the next subsection, we will talk about the role of DOIs (Digital object identifiers). Then, we presented the two major steps of our implementation in the form of article extraction and then citation extraction.

At the end of this section, we have also presented the complete work in the format of an algorithm and flow chart. The next section talks about the main component of citation analysis, in which the following steps were performed: citation analysis on the extracted data, filtration on the extracted data. Then, we discussed the various novel components of our work. The primary questions answered in this section are:

- How different entities can be represented in the form of different qualifiers?
- How different indexing databases are maintaining the concept of uniqueness among these entities?
- What is the availability of the DOIs at author, organization and journal level?
- How was the article data extracted and filtration applied?
- How was the citation data extracted and filtration applied?
- How is the integrity of data maintained among three entities?
- How the citation analysis was performed?
- How we have calculated unified informetrics?

**Chapter 5:** This chapter deals with the statistical analysis of Conflate. Conflate is the result of the complete work. We have categorized the complete results into three sections as follows, author level, organization level, and journal level. We

have also tried to present the complete statistical analysis in the form of tables and elaborated figures. The primary questions answered in this chapter are:

- What is the performance of authors, organizations and journals in terms of number of publications, citations and  $h$ -index?
- What is the difference between self citations, repeated citations, and total citations of authors, organizations and journals?
- What is the performance of journals in terms of impact factor?

**Chapter 6:** This chapter deals with the distributed ledger technology (DLT) based implementation of Conflate. In this chapter, we have explained the concept of analogy and the consensus mechanism of distributed ledger technology. The primary questions answered in this chapter are:

- How is distributed ledger technology mapped with research publications?
- How can “Proof of bibliometric indicators” work as a consensus mechanism?
- How is distributed ledger technology used for the representation of bibliometrics?

**Chapter 7:** The last chapter deals with the summary of the complete work with concluding remarks. In the beginning, we talked about how the complete work is divided and achieved in different steps. In the next section, we have tried to identify the future directions in terms of:

- Different indexing databases,
- Different bibliometric indicators, and
- Different technologies.

In the last section, we have elaborated on how we can extend this work further to a higher level. Here we have also talked about the limitations of our work.

---

---

# CHAPTER 2

---

## Review of literature

In this chapter, literature available on informetrics, indexing databases, and distributed ledger technology is discussed. To explore further, studies related to the comparative analysis of various indexing databases such as Scopus, Web of Science, and Google Scholar, etc., and informetrics such as *h*-index, and impact factor etc., are analyzed. The question of interest is how various authors have studied and presented their work in the context of various indexing databases and how the identified problems in their studies are answered. Further, an identified problem with research gaps is discussed and a possible solution is explored. At the end, the role of distributed ledger technology in various industries and trends is discussed.

### 2.1 Indexing databases

Scientific contributions serve as a catalyst for the advancement of science and society [15]. Citations provide a quantitative evaluation that aids in describing publication trends, research quantity, quality, and author influence to measure the impact of such contributions [16, 17]. Indexing databases such as Scopus, Web of Science (WoS), etc., gather the citations obtained by the papers indexed by them [18, 19].



Hence, generated bibliometrics have a significant impact on the citation data. As a result, variation in the bibliometrics for the same author, based on number of publications, number of citations, and *h*-index, might be found differently in bibliographic databases.

Indexing databases are presented as a data citation road map for scientific publications. In accordance, harmonization and recommendations of the research community, stakeholders, and policy makers with various indexed journals, these indexing databases are the initiatives towards the “life cycle of a research paper”. Transparent data models, robust archiving, and primary sources of research data recommend these indexing databases as authoritative and digital data sources for establishing the data citations [20]. The Web of Science and Scopus are two of the most well-known databases in the world. Web of Science integrated with Science Citation Index (SCI), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI) was introduced in 1997 as a premier resource in the field of study. Later on, a few new citation indexes such as Emerging Sources Citation Index (ESCI), Conference Proceedings Citation Index (CPCI), Book Citation Index (BKCI), Current Chemical Reactions, and Index Chemicus were added, and the Web of Science was renamed to Web of Science Core Collection. In 2004, Elsevier’s Scopus was introduced as an academic bibliographic data source and a powerful competitor to the Web of Science. Scopus claims to cover over 76 million records [21], while WoS claims to cover over 75 million records in its main collection [22].

In 2004, another indexing database named Google Scholar was launched as a massive database of scholarly literature. Google Scholar indexes any academic document on the web by going behind the payment firewalls to provide a comprehensive and multidisciplinary web crawling infrastructure. As compared to other indexing databases, Google Scholar is free to access [23]. Although Google Scholar does not provide official coverage numbers, independent studies have suggested that it covers well over 300 million records [24]. Hence, Google Scholar is considered the most comprehensive coverage database. However, because of the low quality of the metadata

available in Google Scholar and the difficulty in extracting the required credentials, using Google Scholar data in bibliometric analyses is quite challenging [25].

## 2.2 Studies related to comparative analysis of indexing databases

The rapid expansion of bibliographic databases has opened up new possibilities. The authors were able to perform investigations from various viewpoints and with critical relevance due to the multidisciplinary nature and empirical properties of indexing databases. Bibliographic databases are utilised by bibliometricians all around the world to generate comparison statistics [26]. Various stakeholders, such as funding agencies, government organizations, promotion committees, ranking agencies, accrediting agencies, and other stakeholders, use comparative statistics provided by various authors to assess the quality and influence of scientific work. As a result, bibliometric analysis has emerged as a powerful tool in the research publication industry. Different authors have compared Scopus, WoS, PubMed, Google Scholar, and other indexing databases in the literature, based on:

- Availability of digital object identifiers [27]

The author has shed light on the usage of digital object identifiers in the Scopus and Web of Science, based on the data available from 2014-2020. The author has also encouraged scientists to review the use of digital object identifiers in favored publication channels.

- Bibliometric analysis [28]

The author has performed the bibliometric analysis based on the articles cited in Scopus, Web of Science, and Google Scholar for 30 colleges of nursing faculty. An author has concluded that Scopus, Web of Science, and Google Scholar have provided different bibliometrics for an author.

- Coverage [29]

The author has accessed the coverage of the scientific literature available in Scopus and Web of Science based on the academic communities of Norway. Results show that Scopus has comprehensive coverage as compared to the Web of Science.

- Features and citation properties [30]

The author explores the features and citation properties of Scopus and the Web of Science in comparison to Dimensions, a free scholarly database. Results show that Dimensions may be used as a plausible alternative to Scopus and the Web of Science for general citation analysis and citation data.

- Strengths and weaknesses [31]

The author has tried to analyse the relative strengths and weaknesses of Google Scholar, Scopus, and Web of Science. Results suggest that scientists face a trade-off when using different databases; it is up to researchers to use traditional databases or curated databases for diverse coverage.

- Content comprehensiveness and searching capabilities [32]

The author has compared and presented the contrast between Scopus and the Web of Science based on content comprehensiveness, search retrievals, citation counts, and publication counts. The author has concluded that both databases are easy to use and provide useful and informative help in several formats of bibliometrics.

- Assessment of research fields [33]

The author has employed a systematic assessment and in-depth analysis of the Scopus and Web of Science data available for Slovenia between 1996-2011. Results show that there is a difference in the published documents and citations across all major research fields.

- Empirical analysis and classification [34]

The author has provided a preliminary analysis and classification of similarities and differences between Scopus and the Web of Science. Results reveal that both indexing databases have a corpus of errors, but Scopus has more accuracy as compared to Web of Science.

- Journal coverage [35]

The author has collected the bibliographic data for the period between 2013-2019 for the thorough comparison of the same journal articles indexed in Scopus and Web of Science. Results show that there are discrepancies in the number of records based on the journal publisher.

- Citation analysis [36]

The author has presented a case study based on the items published on Scopus, the Web of Science, and Google Scholar. The study concludes that a single indexing database should not be used alone for locating citations, and Scopus and Google Scholar have more comprehensive coverage as compared to the Web of Science and the selection of indexing databases.

- Retroactive comparison of institutions [37]

The author has analyzed the annual number of documents published by Russian universities from 2000-2016 based on the data available in Scopus and Web of Science. Results reveal that the publication count is strongly dependent on the indexing database used.

- Citation counting, citation ranking and  $h$ -index of authors [38]

Author has presented the differences between Scopus and Web of Science based on citation counting, citation ranking and  $h$ -index of authors. Author has concluded that researchers should manually calculate their  $h$  scores instead of relying on automatic systems of indexing databases.

- Rankings of countries [39]

Author has addressed the robustness of country by country ranking based on number of publications and citations derived from Scopus and Web of Science. The study has also discussed the implications for the construction of bibliometric indicators.

## **2.3 Revolution of distributed ledger technology**

Ledgers have been at the heart of trade since the dawn of time, and they are used to record a wide range of transactions, the majority of which involve personal property and currency. They were encouraged to write on clay tablets, which were then transferred to paper, vellum, and papyrus. These ledgers have proven to be crucial in the government's numerous efforts. Ledgers are also highly useful in government projects, according to their potential. [40].

However, computerization, which began as a conversion from manuscript to bytes, is the only notable innovation [41]. Systems assist in the construction of digital ledgers with features and capacities that go beyond traditional manual ledgers. A distributed ledger is essentially a data bank that may be accessed from multiple layouts or groups in a system. Each member of a linkage may have their own personal copy of the ledger. Any change to the ledger is instantly reflected everywhere. The entities can be monetary, legal, or physical [42].

A continuous chain of transactions can be described as distributed ledger technology (DLT). One can track those transactions using their signatures, if the system requires it [43]. During the installation of distributed ledger technology, features such as process validation, authentication, and appropriate processes can be achieved. Distributed ledger technology generates ledgers that can be used in different locations, making them more powerful. Cryptographic keys preserve security and other qualities connected to the accuracy of these ledgers, making the notion more solid in terms of its features.

### **Insight details of Distributed Ledger Technology**

- **Ledger:** A distributed ledger can be thought of as a repository that everyone who contributes to the ledger can access. It is accomplished by the exclusive allocation of records and is dependent on each node in the system.
- **Block:** One portion of the register can be thought of as a block. There is no way to go back or undo the written process once something is written on the register. All segments established in the register will be accompanied by time-stamp signatures, preventing the ledger from being tempered or altered.
- **Record:** A record can be thought of as a grouping of transactions. Initially, every transaction in the distributed ledger is subject to the privilege of record. It then passes on to the next block, and finally to the ledger.
- **Transaction:** The actual event in the system is called a transaction. In order to maintain this essential consistency, every transaction is linked to the one before it. At any point in time, one can retrace the entire transaction record.

## **2.4 Distributed ledger technology based applications**

A plethora of blockchain-based applications provides an opportunity to explore the technology's potential and fully utilize its characteristics. The usage of distributed ledger technology in the research publication sector is a difficult technique to decipher. Although there have been a few other applications of distributed ledger technology in the academic publication sector, there is still a lot of room for incorporating citation analysis with distributed ledger technology. Using distributed ledger technology to systematically build a permanent system in the research publishing sector is a viable option in the interest of stakeholders in the industry. The proposed system would assist its users in achieving a long-term, integrated, and centralized solution for an

author's, organization's, and journal's unified informetrics. The proposed implementation will aid in the calculation of informetrics, assist stakeholders in the examination of various indexing databases, and improve traditional citation calculation trends [44–47].

## 2.5 Consensus mechanism in distributed ledger technology

In distributed ledger technology, consensus is an acceptable point where every participant in a system must accept the events. According to [48], if the data is acceptable, it may be entered into the ledger. In a distributed ledger network, reaching an agreement point is critical [49]. As the number of distributed ledger based applications is growing all the time, the need for consensus methods is growing as well [50, 51]. Newly discovered algorithms give new applications with unique qualities and functional abilities. Consensus mechanisms determine the new options for well-optimized solutions that this technology provides. Consensus refers to an agreement that all parties must accept while making a decision [52, 53].

## 2.6 Research gap

Table 2.1 presents the main findings and limitations identified in studies on comparative analysis of various bibliographic databases. The findings show that while a few studies have attempted to provide a partial solution, none of the literature provides a comprehensive or complete approach to overcome bibliographic database restrictions. Because of these constraints, universities, accrediting organizations, ranking agencies, and recruiting agencies require authors to provide publications, citations, and  $h$ -index counts from all bibliographic databases individually during their job applications and assessments. There is no universal platform for recording or computing single informetrics across numerous bibliographic databases. This has prompted the

development of bibliometrics, which allows authors to provide a single count of publications, citations, and *h*-index across many bibliographic databases. Authors should not offer multiple sets of publications, citations, and *h*-index values for bibliographic databases such as Scopus, WoS, etc.



TABLE 2.1: *List of studies included comparative analysis, based on Scopus, Web of Science, Google Scholar etc. (sorted year wise)*

<b>Key idea or concept</b>	<b>Limitations identified</b>	<b>Any solution proposed?</b>	<b>Reference</b>
To compare the major features of the Web of Science, Scopus, and Google Scholar as a citation databases.	Because traditional indexing databases lack suitable subject indexing, an idea that serves as a remedy to citation-based searching for quantitative evaluations is needed.	×	2005 [54]
To compare strength and weaknesses of PubMed, Scopus, Web of Science, and Google Scholar databases.	Databases are compared in terms of content and practical applications. There are no such restrictions listed.	×	2008 [55]
To compare h-indices of highly cited researchers of Israel based on their citation count in Web of Science, Scopus and Google Scholar.	Across databases, disciplinary disparities in coverage and citation numbers are found.	×	2008 [56]
To gauge the comparability in determining the h-index from Scopus and Web of Science for 10 universities.	Both databases have significant disparities in terms of content and referenced sources.	×	2009 [57]

Table 2.1 continued from previous page

Key idea or concept	Limitations identified	Any solution proposed?	Reference
To compare two instruments like Scopus and Web of Science for a typical university in Portugal.	Different abstracting policies are discovered, as well as apparent database construction errors are identified.	×	2009 [58]
Three citation databases are compared with reference to book - introduction to informetrics.	Citations across databases are definitely comparable, according to the findings. However, no single citation database can replace the others.	×	2010 [59]
Three citation resources are compared to find the one with most representative citation coverage.	The findings reveal that the retrieved data varies in terms of citation counts.	×	2013 [60]
Journal coverage across Scopus and Web of Science is described.	The findings suggest that using either of these databases to evaluate research could be biased. As a result, both must be taken with caution.	×	2016 [61]

Table 2.1 continued from previous page

Key idea or concept	Limitations identified	Any solution proposed?	Reference
Systematic and comprehensive comparison of coverage across Scopus, Web of Science and Google Scholar is provided.	For cross-disciplinary comparisons, all three databases are sufficient. However, the results reveal that different measures affect the outcomes in different datasets.	×	2016 [62]
A light has been shed on the availability of DOIs in Scopus and Web of Science in publication items.	Because neither database has complete DOI availability, authors are advised to use DOI based establishments.	×	2016 [63]
Research data of central universities in India are studied for their ranking and policy formulation.	Generalized model is presented to introduce the idea of quality-quantity composite index.	Partial	2016 [64]
Google Scholar, Web of Science, and Scopus are compared based on 252 subject categories.	According to the study, Google Scholar has more citations than Scopus and Web of Science. Google Scholar can be thought of as a super set of Scopus and Web of Science.	Partial	2018 [65]

Table 2.1 continued from previous page

Key idea or concept	Limitations identified	Any solution proposed?	Reference
Bibliometrics based on highly cited documents in Scopus, Web of Science and Google Scholar is explored.	Based on counts of highly referenced papers, the study shows that these databases miss a considerable quantity of information (when compared).	×	2018 [66]
Publications for Jordanian authors are studied based on literature databases such as Scopus, Web of Science, Pubmed etc.	Scopus, Web of Science, Pubmed, and other databases reveal variances in terms of coverage, focus, and tools.	×	2019 [67]
Web of Science and Scopus are compared based on their language coverage of publications.	For both Scopus and Web of Science, the results obtained at the document level, languages, and key areas differ from those found at the journal level.	×	2019 [68]

Table 2.1 continued from previous page

Key idea or concept	Limitations identified	Any solution proposed?	Reference
Retroactive growth, correlation, and coverage of universities is validated based on Scopus, Web of Science and Google Scholar.	In terms of overall number of publications, institutional productivity varies across Google Scholar, Scopus, and Web of Science.	×	2019 [69]
Comparative, dynamic, and empirical study is presented based on academic papers available in Scopus and Web of Science.	A more thorough research based on Scopus and Web of Science content is required.	×	2020 [70]
Citation coverage is presented based on Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations COCI.	The findings show that no single bibliographic database is appropriate. Future research may determine which data source is best suited to stakeholders' needs.	Partial	2021 [71]

Table 2.1 continued from previous page

Key idea or concept	Limitations identified	Any solution proposed?	Reference
Comparative analysis of journal coverage is aimed for Scopus, Web of Science and Dimensions.	The findings show that the journal coverage of databases varies greatly.	×	2021 [72]
Impact of author ranking based on Scopus and Web of Science is introduced with an improvement to <i>h</i> -index.	For low-rank authors, the results show that the <i>h</i> -index calculated in Scopus and Web of Science varied significantly.	Partial	2022 [73]

## 2.7 Discussion and summary

At the beginning of this chapter, we talked about the concept of various indexing databases and related informetrics. A detailed literature review is presented based on two broad categories. The first is indexing databases, and the second is studies related to comparative analysis of indexing databases. Later on, the concept of distributed ledger technology in the context of its revolution, adoption, and consensus mechanisms is discussed. In this revolution, various technical terms associated with distributed ledger technology are discussed, followed by a number of applications introduced by various authors in context with distributed ledger technology based implementation. During the study of such applications, it is observed that in almost every key area such as education, healthcare, supply chain management, e-commerce, exchange of information etc., distributed ledger technology has been introduced. The work mechanisms of various consensus algorithms with their implementations are also explored.

During the literature review of indexing databases, it is identified that there are a number of authors worldwide who have talked about the indexing databases on various parameters. A few authors have also listed the limitations of indexing databases and a few of them have also tried to provide partial solutions to the limitations identified, but none of the studies or authors has tried to provide a comprehensive solution. Hence, based on the literature review, a research gap is presented. In the next chapter, the limitations of  $h$ -index as an important and well known informetrics parameter are explored. At the end a complementary solution, named as  $h_c$  is introduced.

---

---

## CHAPTER 3

---

# Indexing databases and informetrics

In this chapter, the significance of using various indexing databases for the calculation of informetrics is discussed. The question of interest is, how the results generated from the different indexing databases, such as Scopus, and Web of Science etc., result in different informetrics for the same author. At the end, results are analyzed and a positive impact in context of  $h$ -index was observed specifically for low ranked authors. Moreover, a new complementary index named as  $h_c$  is also presented as a supplementary approach to the  $h$ -index.

### 3.1 Impact of informetrics on authors

The number of publications, citations, and contribution of an author to scientific knowledge and society are the finest criteria for scientific appraisal [74]. Citation analysis, on the other hand, is an important tool in scientometrics since it is used to evaluate an individual's research performance in the academic community [75–78]. In addition to the citations, the number of publications and the  $h$ -index have



high validation in research evaluation. The  $h$ -index's popularity and attention among scientists stems from its straightforward calculation. Rather than showing individual variables such as the number of publications, citations, and so on, which provide a single dimension of an author's performance, the  $h$ -index brought multidimensional presentation (quantity and impact), all with a single integer number [79–82].

As a result, it is regarded as a balanced method of combining and evaluating an author's broad scientific contribution [83]. According to [84], an author's  $h$ -index is equivalent to a journal's impact factor. Authors [85] have also identified the  $h$ -index as a measure of journal credibility and evaluations. Because of its popularity, various indexing databases like as Scopus, WoS, and others publish an author's calculated  $h$ -index on their websites [86, 87]. Thus, in order to conduct a fair appraisal of an individual within a university/institution, funding agency, scientific community, and so on, the evaluated scientometric parameters must be field, discipline, and time normalized [88].

For valid  $h$ -index assessments, indexing databases' coverage, consistency of data, saving choices, data fields, browsing options, searching options, analytical tools, exporting options, and data accuracy should all be carefully assessed. As proposed by [89], “*A scientist has index  $h$  if  $h$  of his/her  $N_p$  papers have at least  $h$  citations each and the other  $(N_p - h)$  papers have  $\leq h$  citations each.*” In bibliometrics, the  $h$ -index is one of the most essential, robust, primitive, quantitative, and singular measures used to assess the quality, impact, influence, and relevance of an individual's work.

Researchers have demonstrated the use and usefulness of the  $h$ -index in determining author rankings, university rankings, journal impact, and so on. A few studies have examined the authors' bias and performance across a wide variety of citations but found no significant differences between the globalized and average citation variations [90]. In the literature, various ways of analyzing the author's ranking have been utilized. Authors have also demonstrated how they used the PageRank algorithm to rank the author co-citations network [91–94]. Usman et al. used several assessment

characteristics such as  $h$ -index, citations, publications, authors per paper,  $g$ -index,  $hg$ -index, R-index, e-index,  $h'$ -index, w-index, and others to rank authors in their study [95–100].

Hirsch mentions some disadvantages in his core paper in addition to the numerous benefits [101]. The most widely mentioned negative is that it does not account for the influence of highly cited papers, which means it undervalues scientists' academic success. Many new indices have been proposed in this line to circumvent this issue, one of which is the  $g$ -index [102–106]. The limitations of not considering highly cited papers, have encouraged Leo Egghe to propose  $g$ -index in 2006 as follows: “*A set of papers has a  $g$ -index  $g$  if  $g$  is the highest rank such that the top  $g$  papers have, together, at least  $g^2$  citations. This also means that the top  $g + 1$  papers have less than  $(g + 1)^2$  papers*” [107].

By assigning more credit to highly cited papers and having more discriminatory capacity to represent an author's scientific contribution, the  $g$ -index outperforms the  $h$ -index [108, 109]. Leo Egghe also proposed the addition of fictional articles with no citations to overcome the constraints and complete the  $g$ -index computation. The  $g$ -index takes care of highly cited publications, however it considerably diminishes the impact of the most cited paper [110].

The current study tackles this difficulty by creating  $h_c$ , a supplementary index that complements the  $h$ -index by including the weight of the most cited publication while maintaining the  $h$ -index's most important benefits. Here, a complementary analysis to the existing  $h$ -index, named  $h_c$  (see Algorithm 1) is proposed. The impact of the most cited paper is computed and it's added to the  $h$ -index of an author. Equation 3.1 is used to get the impact of the highest cited paper as

$$h^k < H_{cite}, \quad (3.1)$$

where  $h \geq 0$  is the  $h$ -index.  $H_{cite} \geq 0$  is the citation count of an author's highest cited paper,  $k$  is the weight of the highest cited paper and  $k \leq h$ . Now  $h_c$  is computed as

$$h_c = h + k. \quad (3.2)$$

$h_c \geq h$  is the compliment of  $h$ .

---

**Algorithm 1** Calculation of  $h_c$

---

```

set  $k = 0, i = 0$ 
while  $k \leq h$  do
  if  $h^k < H_{cite}$  then
     $i = k$ 
     $k = k + 1$ 
  else
    break
  end if
end while
 $h_c = h + i$ 

```

---

**Example I:** Let's say an author is having  $n$  publications with  $h = 0$  and  $H_{cite} = 0$ . In this case, none of the paper got citations, hence  $h = 0$ . Using 3.1 and 3.2, the value of  $h_c$  will be 0.

**Example II:** Let's say an author is having  $n$  publications with  $h = 1$  and  $H_{cite} = 1$ . In this case, atleast one paper got cited once. Using 3.1 and 3.2, the value of  $h_c$  will be 1.

**Example III:** Let's say an author is having  $n$  publications with  $h = 1$  and  $H_{cite} = 2$ . In this case, only paper got 2 citations. Using 3.1 and 3.2, the value of  $h_c$  will be 2.

**Example IV:** Let's say an author is having  $n$  publications with  $h = 2$  and  $H_{cite} = 2$ . In this case, two or more papers got cited twice each and highest citation is 2. Using 3.1 and 3.2, the value of  $h_c$  will be 3.

**Example V:** Let's say an author is having  $n$  publications with  $h = 1$  and  $H_{cite} = 3$ . In this case, one paper got cited with 3 citations. Using 3.1 and 3.2, the value of  $h_c$  will be 2.

Table 3.1 explains the above mentioned examples representing different scenarios of the research productivity.

TABLE 3.1: *Demonstration of  $h_c$ .*

Example	$H_{cite}$	$h$	$h^k < H_{cite} \ \& \ k \leq h$	$k$	$h_c$
I	0	0	$0^0 < 0 = \text{F}$	0	0
II	1	1	$1^0 < 1 = \text{F}$	0	1
III	2	1	$1^0 < 2 = \text{T}$	1	2
			$1^1 < 2 = \text{T}$		
IV	2	2	$2^0 < 2 = \text{T}$	1	3
			$2^1 < 2 = \text{F}$		
V	3	1	$1^0 < 3 = \text{T}$	1	2
			$1^1 < 3 = \text{T}$		

## 3.2 Author's ranking in Scopus and WoS based on informetrics

For both Scopus and WoS, Figure 3.1 illustrates the ranking of authors for five fields in terms of  $h$ ,  $h_c$ , and  $g$ -indices. The ranks are ordered in descending order by Scopus  $h$ -index, and the  $g$  and  $h_c$ -indices of the authors are presented separately. In all disciplines, there is a lot of variation for writers with different  $h$ -indexes.

At the tail, the variations in  $h_c$  with regard to  $h$  are more pronounced. The probability density function of  $h$  and  $h_c$  is shown in the inset. The influence of  $h_c$  on lower-ranked authors may be seen in the tiny shift of the  $h_c$  to the right. The most significant influence is on *Social Sciences*, where  $k = 2$  increases the index value of 34.1% of authors in Scopus and 40.0% in WoS. Similarly, for  $k = 3$ , negligible increase is recorded.

*Health and Medical Sciences* is the second highest, with 32.1% in Scopus and 27.4% in WoS for  $k = 2$ . In this discipline,  $k = 3$  has a negligible effect. In *Biochemistry and Molecular Biology*, the increase is 22.7% in Scopus and 23.9% in WoS for  $k = 2$ , whereas the impact is minimal for  $k = 3$ .

For  $k = 2$ , the total impact is 25.0% in Scopus and 25.6% in WoS, and for  $k = 3$ , the overall impact is 1.1% in Scopus and 0.6% in WoS. Scopus and WoS have nearly identical overall impact; nevertheless, there are differences in fields. WoS provides a consistent ranking in *Health and Medical Sciences*, whereas Scopus gives a stable ranking in *Natural Sciences* and *Social Sciences*. The difference is insignificant in *Biochemistry and Molecular Biology* and *Engineering*.

### 3.3 Comparative analysis of informetrics

For both Scopus and WoS, Figure 3.2 shows the comparison between (a)  $h$  and  $h_c$ , and (b)  $h$  and  $g$ . On various  $h$ -indexes, a correlation coefficient is calculated. Figure 3.2(a) captures the fluctuations for lower ranking authors, i.e. for  $h \leq 10$  with mean correlation 0.9 for both Scopus and WoS. (a). The mean correlation for other cut-offs is above 0.95, indicating minor volatility. Figure 3.2(b) shows that the fluctuations are higher with varying  $h$ -index. The minimum index value has been increased or maintained in all disciplines, i.e.  $h_c \geq h$ , as indicated in Table 3.2. In some circumstances,  $h_c$  outperforms  $g$  in terms of minimum index value. For both Scopus and WoS, there is no change in the maximum index value, i.e.  $h == h_c$ , and a little variation in median values. For all fields, the average index value is nearly the same across Scopus and WoS.

TABLE 3.2: Statistics of  $h$ ,  $h_c$  and  $g$ .

Disciplines	IDX	Min			Max			Median			Average			SD		
		$h$	$h_c$	$g$	$h$	$h_c$	$g$	$h$	$h_c$	$g$	$h$	$h_c$	$g$	$h$	$h_c$	$g$
Biochemistry and Molecular Biology	S	4	7	11	79	80	137	22	23	43	25.2	26.4	47.5	15.3	15.2	27.1
	W	5	7	10	77	78	133	22	23	41	24.5	25.7	46.2	15.0	14.9	27.3
Engineering	S	2	4	3	64	65	102	18	19	31	20.7	21.9	36.2	14.0	13.7	23.0
	W	1	2	2	62	63	99	16	17.5	30	19.6	20.8	34.4	13.5	13.4	22.8
Health and Medical Sciences	S	2	4	4	91	92	173	17	18	33	21.6	23.0	41.9	16.1	16.0	33.9
	W	2	4	3	95	96	168	16	17	30	20.6	21.9	39.6	16.1	16.0	33.9
Natural Sciences	S	4	5	6	50	51	98	18	19.5	36	21.2	22.4	38.1	12.2	12.1	22.1
	W	2	4	4	49	50	101	17	18	33	20.6	21.8	36.8	12.3	12.2	22.0
Social Sciences	S	1	2	2	72	73	146	13	15	25	17.0	18.4	31.6	13.9	13.7	26.7
	W	1	1	1	68	69	141	11	13	23	15.4	16.8	28.8	13.2	13.1	24.9

Figure 3.3 depicts the  $h_c$  growth curve depending on  $H_{cite}$ . The respective  $h_c$  is determined for different values of most cited paper ( $H_{cite}$ ). We kept the value of

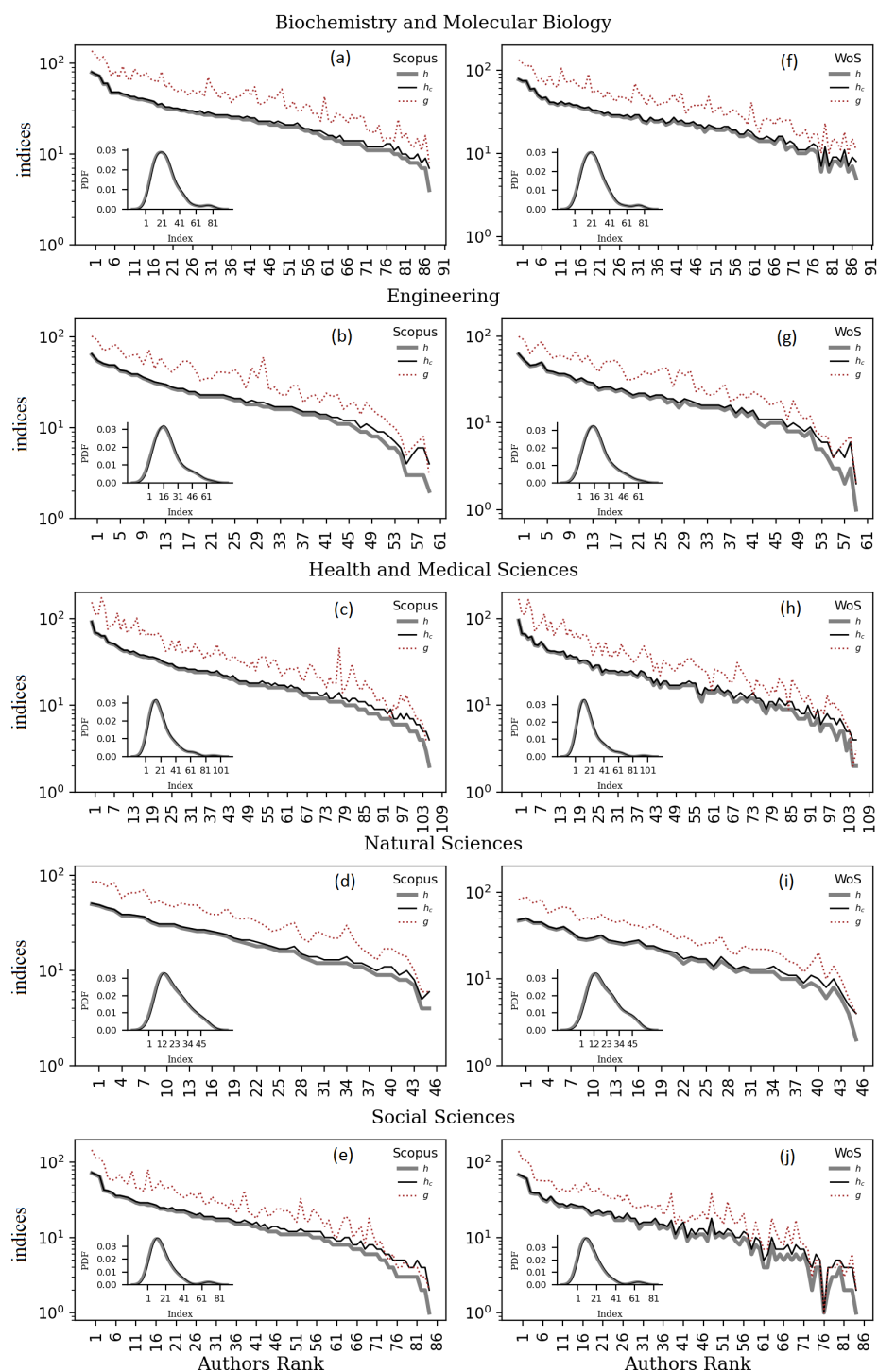


FIGURE 3.1: Author ranking using the  $h$ ,  $g$ , and  $h_c$  indices.

$h$  at 2 in Figure 3.3. For any value of  $h$ , a comparable growth curve can be created. The impact is plainly seen in the early stages of the growth curve, but it fades after that. We can observe that the  $h_c$  makes a difference for an author with  $h = 2$  and

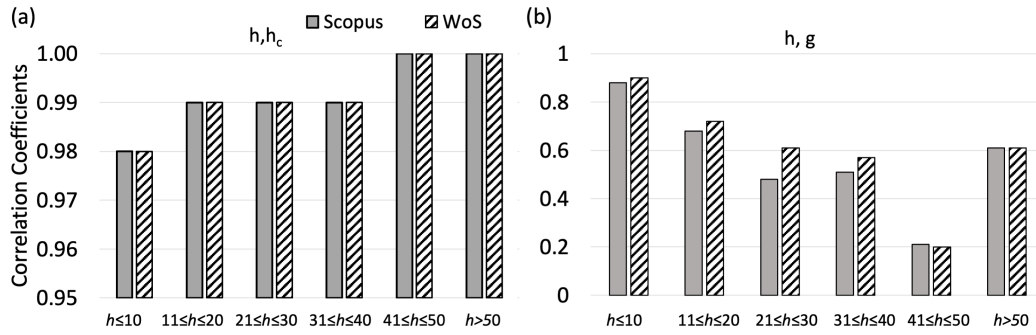
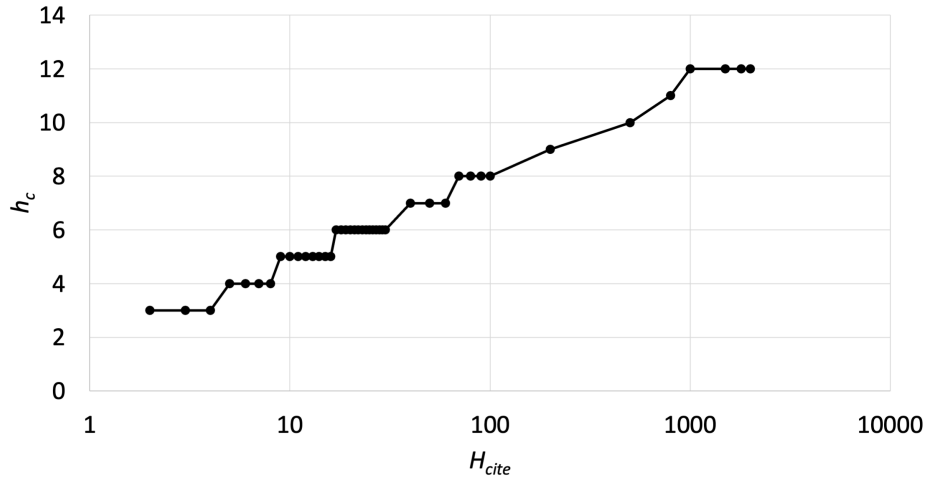


FIGURE 3.2: Correlation coefficients of (a)  $h$  and  $h_c$ , and (b)  $h$  and  $g$  for both Scopus and WoS.

$8 < H_{cite} < 100$ , i.e., the author's most cited article is given more weight.



Scopus, on the other hand, displays 34.5% and WoS 32.2% of authors in the range  $11 \leq h \leq 20$ , respectively, although this count is enhanced for  $h_c$  to 36.6% for Scopus and 34.8% for WoS. As a result, the number of authors in Scopus and WoS has increased by 2%. Scopus shows 4.4% and WoS shows 4.2% of authors for higher-ranked authors ( $h > 50$ ), while  $h_c$  indicates no impact on authors ranking at a higher level.

For both Scopus and WoS, Table 3.3 shows the distribution of authors (in %) depending on  $h$  and  $h_c$ . For all fields, the fraction of authors with  $h \leq 10$  is higher in WoS. For authors with  $h \leq 10$  in Scopus and WoS, *Social Sciences* has the highest count (35.3%) while *Biochemistry and Molecular Biology* has the lowest count (11.4%). Furthermore, in *Health and Medical Sciences* and *Natural Sciences*, where the authors ranked  $h \leq 10$ , a 6% change in ranking from  $h$  to  $h_c$  is observed. The change in the remaining disciplines is between 2-5% approximately.

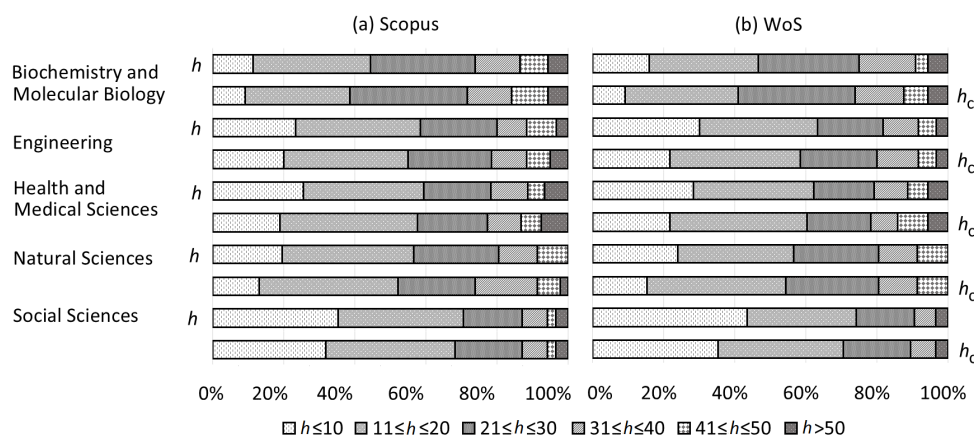


FIGURE 3.4: For (a) Scopus and (b) WoS, the fraction of authors,  $h$  and  $h_c$  wise.

### 3.5 Discussion and summary

Based on informetrics calculated from indexing databases, an impact on an author's ranking was explored. The pioneering and groundbreaking work of J. E. Hirsch was also studied by focusing on one of the limitations of  $h$ -index [111]. The  $h$ -index considers both the quantity and impact of publications, but ignores the influence of highly cited papers, which undervalues the effort. Following the establishment of the



TABLE 3.3: Proportion of authors with varying  $h$  and  $h_c$  for Scopus ( $S$ ) and WoS ( $W$ ).

Disciplines	ID	No. of Authors [%]											
		$h \leq 10$		$11 \leq h \leq 20$		$21 \leq h \leq 30$		$31 \leq h \leq 40$		$41 \leq h \leq 50$		$h > 50$	
		$h$	$h_c$	$h$	$h_c$	$h$	$h_c$	$h$	$h_c$	$h$	$h_c$	$h$	$h_c$
Biochemistry and Molecular Biology	S	11.4	9.1	33.0	29.5	29.5	33.0	12.5	12.5	8.0	10.2	5.7	5.7
	W	15.9	9.1	30.7	31.8	28.4	33.0	15.9	13.6	3.4	6.8	5.7	5.7
Engineering	S	23.3	20.0	35.0	35.0	21.7	23.3	8.3	10.0	8.3	6.7	3.3	5.0
	W	30.0	21.7	33.3	36.7	18.3	21.7	10.0	11.7	5.0	5.0	3.3	3.3
Health and Medical Sciences	S	25.5	18.9	34.0	38.7	18.9	19.8	10.4	9.4	4.7	5.7	6.6	7.5
	W	28.3	21.7	34.0	38.7	17.0	17.9	9.4	7.5	5.7	8.5	5.7	5.7
Natural Sciences	S	19.6	13.0	37.0	39.1	23.9	21.7	10.9	17.4	8.7	6.5	0.0	2.2
	W	23.9	15.2	32.6	39.1	23.9	26.1	10.9	10.9	8.7	8.7	0.0	0.0
Social Sciences	S	35.3	31.8	35.3	36.5	16.5	18.8	7.1	7.1	2.4	2.4	3.5	3.5
	W	43.5	35.3	30.6	35.3	16.5	18.8	5.9	7.1	0.0	0.0	3.5	3.5

$h$ -index, scientists offered numerous  $h$ -index versions in order to better an individual's study evaluation [112]. Some, such as the  $g$ -index, have gained significance; yet, each index is deficient in some way.  $h_c$  is presented, as a supplementary approach to the  $h$ -index, which is based on the  $h$ -index. The new index  $h_c$  is based on the same ranking, and we discovered that substantial fluctuations arose for authors rated  $h \leq 10$  in both Scopus and WoS; however, the variation/fluctuation in WoS is bigger than in Scopus. Scopus and WoS produce different results when analysing disciplines.

Because of its simplicity, and in addition to the  $h$ -index,  $h_c$  could provide important insight into youthful or lower-ranked authors, thereby improving an individual's rating within a discipline. It also emphasizes the value of an individual's work by taking into consideration the  $h$ -index as well as the contribution of the most-cited piece of an author. In the next chapter, a unique approach named unified informetrics is introduced as a novel solution for the identified research gap.

---

---

## CHAPTER 4

---

# Linking of indexing databases and generation of unified informetrics (UI)

In this chapter, a discussion is presented based on the different kinds of entities used to examine the depth of the problem. How is the data for these entities fetched and filtered? The question of interest is how the fetched data is mapped and how the uniqueness among different entities is maintained. So, a mechanism is derived to map both the concepts together and maintain the uniqueness among entities. Further, an algorithm is proposed to perform the extraction of data and assign weight to the informetrics for the generation of unified informetrics.

## 4.1 Entity specification and linkage of citation analysis

In 2005, Hirsch introduced *h*-index as a simple, straight forward and significant striking indicator in terms of publications and citations to measure the scientific output of an individual. He defined the mechanism as “*A scientist has index  $h$  if  $h$  of his/her  $N$  papers have at least  $h$  citations each and the other papers have no more than  $h$  citations each*”. Due to the simplicity of the *h*-index, it has achieved high success among its stakeholders. This novel dimension of measuring the scientific output of an individual has given a new meaning to the publication industry. Different stakeholders, like government organizations, accreditation agencies, ranking agencies, and funding agencies, have recognized this as a considerable factor for the measurement of a scientific contribution.

Individuals publish their scientific work in different journals which are indexed in different indexing databases. Scopus (<https://www.scopus.com/home.uri>), Web of Science (<https://www.webofknowledge.com/>), Microsoft Academic (<https://academic.microsoft.com/home>), Google Scholar (<https://scholar.google.com/>), OpenAIRE (<https://www.openaire.eu/>), Mendeley (<https://www.mendeley.com/>), PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) and Zenodo (<https://zenodo.org/>) are the recognized sources from where one can access the scientific work of different authors, organizations and journals.

These indexing databases are used as the primary resources for the calculation of the *h*-index of an author, organization, and journal. As all these indexing databases use their own concepts for the recording of publications and calculation of citation counts of an author, organization and journal, one is liable to get the same as well as different publication count and citation count in these indexing databases. In context of which, an individual will see different *h*-index values in these indexing databases. Further, as the calculation of *h*-index is a simple and straight-forward formula, it does not consider the repetition of publications indexed in various indexing databases, nor

does it counter the citations which are received by these repeated publications in multiple indexing databases.

There is an essential need of a platform where multiple indexing databases can be combined so that an individual can get a single publication count and single citation count for his scientific work [113]. Furthermore, there should be an index that can assign weight to the publications and citations received in multiple indexing databases for an individual. This can help an individual to judge his scientific work across all indexing databases, which can certainly help him to calculate a single index for his scientific work.

#### **4.1.1 Author level bibliometrics**

An author is considered a person who conceives the idea of scientific work. An author dreams of the idea so that it can become a reality for others. In entity specification, we have categorized an author as the first component of a system. An author is responsible for deciding the life cycle of his idea until it is delivered to its stakeholders. An author can be linked or associated with an organization. It can be an academic organization, it can be any profit or non-profit organization or it can be a government, private or public organization. An author who is interested in writing any scientific work may have certain associations with publication organizations as well.

Authors do have certain specialized areas or domains about which they always talk and always write. These days, we can see authors collaborating at a national and international level as well. Such collaborations with multiple authors give a new meaning to the scientific work done by such individuals. Scientific work done in such collaborations expects a high level of quality of work among individuals [114]. The primary reason for considering authors as the first and most important entity in our work is the fact that authors are the starting point of each scientific work in one or another manner. Different indexing databases use their own concept of keeping a record of authors. A few indexing databases maintain the author profiles with their

names, and a few others maintain the author profiles with author ids. In Scopus, we can find that author profiles are identified by author ID and they may contain the author's first name, last name, affiliation and other related information such as publications and citations. On the Web of Science, we can find that author profiles are identified by researcher ID and they may contain affiliation details, research field details, the name of an author and other related information such as publications and citations.

### **4.1.2 Organization level bibliometrics**

An author is considered as a person who conceives an idea of scientific work. An organization can be considered as an entity that nurtures an author. It is just like an ideal example of parenting, which helps an author to grow, believe, and cherish his ideas in society. Organizations always support, motivate, and encourage authors so that they can develop, cultivate, and sustain the growth of their scientific work.

Primarily, authors who are associated with scientific work are affiliated with one or more academic organizations. These organizations are generally considered for higher education and can be categorized as state universities, deemed universities, central universities, private universities, IIMs, IITs, IIITs, IISCs, IISERs and NITs etc. All of these organizations have certain specializations like engineering, management, pharmacy, medical, law, architecture, and dentistry as well [115].

The primary reason for considering organizations as a second and important entity in our work is the fact that organizations are responsible for the growth of an author and their scientific work in one or another manner. Different indexing databases use their own concept of keeping records of organizations. Few indexing databases keep organization profiles with their names, while others keep organization profiles with organization ids. In Scopus, we can find that organization profiles are identified by organization ID and they may contain information such as organization name, organization address, and other related information such as publications and citations. On the Web of Science, we can find that organization profiles are identified

by their names and they may contain address details, the name of an organization and other related information like publications and citations.

### **4.1.3 Journal level bibliometrics**

Journals are the final destinations of the authors. It is just like parents marrying their daughter and sending her to the outer world to start a new journey in her life. Authors send their scientific work to the journals so that it can start the journey of its life in terms of citations. Authors have to wisely select the journals before sending their scientific work to them. Quality scientific work published in an appropriate journal can result in very high citations.

Journals do not accept all scientific articles. Acceptance is subject to different subject areas like agricultural and biological sciences, arts and humanities, biochemistry, genetics and molecular biology, business, management and accounting, chemistry, computer science, decision sciences, dentistry, earth and planetary sciences, economics, econometrics and finance, energy, engineering, environmental science, health professions, immunology and microbiology, materials science, mathematics, medicine, multidisciplinary, neuroscience, nursing, pharmacology, toxicology and pharmaceuticals, physics and astronomy, psychology, social sciences, and veterinary etc. [116].

The primary reason for considering journals as a third and important entity in our work is the fact that journals are responsible for the growth of the scientific work of an author. The international standard serial number (ISSN) is a way for different indexing databases to keep track of journals. In Scopus and Web of Science, we can find that journals are identified by their ISSN numbers and names, and they may contain other related information or bibliometric indicators linked with the journal as well.

### **4.1.4 Entity identifiers**

Different indexing databases use their own concept of maintaining the records of different entities. In one database, entities may be distinguished by name, and in

another by ID. For example, Scopus uses the concept of *author ID* to uniquely identify author information in its database, whereas the Web of Science uses the concept of *Researcher ID*. However, indexing databases have their own pattern of keeping a record of author information. So, if one has to publish work in these databases, he/she must keep a record of all such pointers. As the identified problem is dependent on these pointers, there should be a common platform where all such pointers may be observed under one umbrella [117]. The possible solution available for this problem is *Orcid ID*. It's a persistent digital identifier that helps to link complete research work with a single ID. This ID can be used as a primary key to bind the different entities of authors across multiple indexing databases.

For organizations, Scopus uses the *affiliation ID* as a unique ID. With the help of this ID, the name of affiliation may be queried and a database of the same may be maintained. On the Web of Science, we do not find the concept of IDs for organizations but it has its own way of storing the organization information in the database. After retrieving organization names from Scopus, one could either manually map those names with the Web of Science or assign random IDs to these organizations.

Journals are the prime sources of publications. Both Scopus and Web of Science keep the record of journals with the key feature of ISSN (International Standard Serial Number). It is a unique number assigned to each and every journal for its entry in the database. Usually, journals have their own subject areas associated with them. Every journal is a specialized version of its subject and related fields. To combine journals across multiple databases, we can utilize the ISSN number, which may act as a primary key.

#### **4.1.5 Digital object identifiers**

The digital object identifier (DOI), also known as a URL (Uniform Resource Locator), is a generic standard for identifying many sorts of items or metadata on the internet, such as documents, photographs, and audio files. It is intended to provide a reliable linking alternative for sharing actionable identification with interested people

or the community [118]. Permanent item identity and uniqueness, interoperability, permanence, and network accessibility are all key advantages in diverse applications. DOI has been used to convey metadata in both physical and electronic formats in digital settings since 2000. The DOI remains fixed during its lifespan, but metadata may change over time. As a result, the DOI name can be as long as it wants.

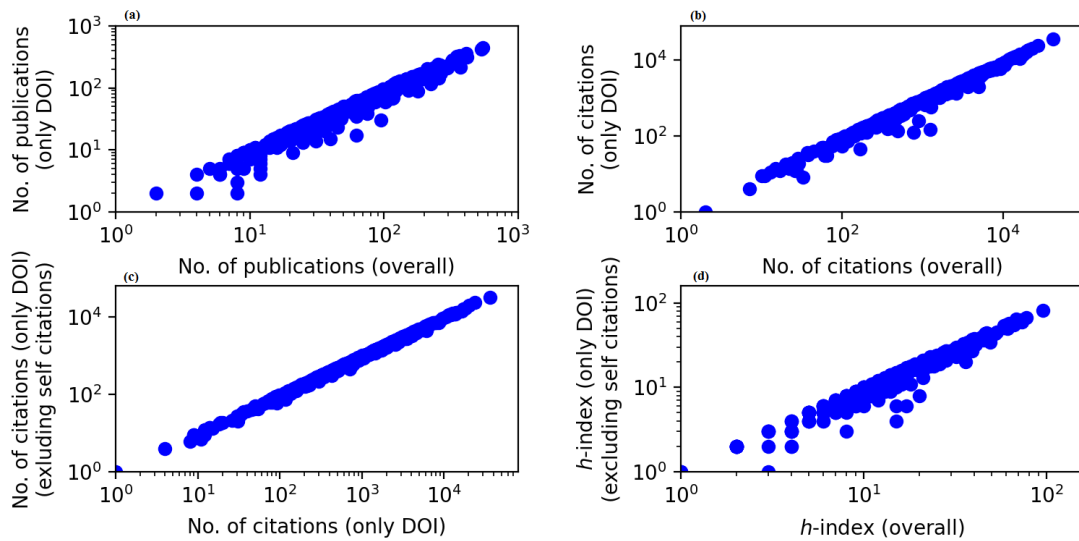
DOI is a character string made up of two parts: a prefix and a suffix, separated by a forward slash '/'. The suffix component implies any user-entered string, whereas the prefix portion denotes a unique naming authority (typically an organization) (usually represents actual identity). The identifier component becomes an actionable link when both components are combined, just like any other URL [119, 120]. DOI is becoming increasingly important in the scientific publishing sector. DOIs have increased in popularity as a global collaboration since DataCite began providing DOIs to scientific papers and research datasets in 2009 [121]. DOIs are used in scientometrics by various indexing databases such as Scopus, Web of Science, Google Scholar, and others to ensure the accuracy of scientific data. They frequently utilize DOIs to reference and share publication data with the scientific community. The availability of DOIs across several indexing databases determines the scientific data's potential stability [122, 123].

Many stakeholders, including academic institutions, research organizations, government entities, promotion committees, and accreditation agencies, are interested in measuring an individual's or a group's research contributions [124]. It could be for individual employment, promotion, tenure, grant release, or literature search etc. Various indexing databases are used by these stakeholders to retrieve real information such as publications, citations, and the *h*-index of an author, organization, or journal [125]. Retrieved informetrics from indexing databases may have various disguised accelerations, such as considering publications and citations without DOI information, and secondly, considering self citations for an undue gain of citations and rise in *h*-index [126–129].



#### 4.1.5.1 Author level bibliometrics

The comparison of (a) documents (with and without DOI), (b) citations (with and without DOI), (c) self citations and citations (with DOI), and (d)  $h$ -index is shown in Fig. 4.1 (with and without DOI and self citations). The analysis was carried out on 400 authors based on comparative analysis of publications, citations, self-citations, and  $h$ -index with and without DOI. According to the study, the overall number of documents has reduced to 26,101 from 31,732, accounting for 82.3% of all documents. The total number of citations for 400 authors is 10,24,808, which is reduced to 8,35,962 when only DOI citations are included, accounting for 81.6% of total citations. Citations per author have declined by 19% on average. According to the initial analysis of self citations, there are 13 authors with no self citations, accounting for 3.3% of total authors, 263 authors with less than 10% self citations, accounting for 65.8% of total authors, 101 authors with less than 20% self citations, accounting for 25.3% of total authors, and 23 authors with more than or equal to 20% self citations, accounting for 5.8% of total authors. If we consider self citations, DOI and non-DOI based documents, the minimum  $h$ -index is 1 and the maximum  $h$ -index is 95. However, if we consider only DOI based documents and exclude self citations, there is no change in the minimum  $h$ -index and a 13 point change in the maximum  $h$ -index, which comes to 82. 70 authors do not observe any change in the  $h$ -index if we follow DOIs and exclude self-citations. With a minimum  $h$ -index of 1 and a maximum  $h$ -index of 23, 314 authors (78.5% of authors) see a decline of 1 to 9 points, with a minimum  $h$ -index of 1 and a maximum  $h$ -index of 64, and an average  $h$ -index of 17.4. With a minimum  $h$ -index of 4, a maximum of 82, and an average  $h$ -index of 36.9, 16 authors out of 400 (4% of authors) had noticed a shift of 10 to 16 points. Table 4.1 shows the results of 400 authors based on five disciplines: *Life Sciences*, *Engineering*, *Sciences*, *Social Sciences*, and *Humanities*. *Sciences* is on top with 88.1% DOI documents, while *Engineering* is at the bottom with 78.8% DOI papers. The *Humanities* category garnered 88.6% of valid DOI citations, followed by *Life Sciences*. *Engineering* has the most self-citations (11.0%), followed by *Sciences*. *Social Sciences* obtained a minimum of 5.3% of all self-citations. *Sciences* has a

FIGURE 4.1: *Comparative analysis of 400 authors based on DOI information.*

$h$ -index of 23.2, which includes self-citations and takes into account all documents, including those with and without DOI. *Life Sciences* has a  $h$ -index of 21.8. After eliminating self-citations and just evaluating DOI-based papers, the average  $h$ -index for *Sciences* is 19.4.

TABLE 4.1: *Comparative analysis of 400 authors based on DOI information.*

Author Disciplines	No. of pubs	(%) of pubs (only DOIs)	No. of cites	(%) of cites (only DOIs)	(%) of self cites (only DOIs)	Avg. $h$ index	Avg. $h$ index (only DOIs, exc. self cites)
Life Sciences	18257	82.2	631244	82.8	7.1	21.8	19.0
Engineering	5658	78.8	138631	73.3	11.0	20.9	16.6
Sciences	3187	88.1	121752	81.5	9.4	23.2	19.4
Social Sciences	3113	80.5	94195	82.6	5.3	13.2	11.6
Humanities	1517	86.9	38986	88.6	8.2	19.8	17.9

#### 4.1.5.2 Organization level bibliometrics

The comparison of (a) documents (with and without DOI), (b) citations (with and without DOI), (c) self citations and citations (with DOI), and (d)  $h$ -index is

shown in Fig. 4.2 (with and without DOI and self citations). The analysis is carried out on 100 organizations, and document differences are observed in each of them, with a minimum difference of 45 documents, a maximum difference of 10,944 documents, and an average difference of 1,893 documents per organization. There was a difference of 5,000 or more documents with DOI and without DOI in 11 organizations, and a drop of less than 100 documents in four organizations. The average number of documents is 7,971.6, versus 6,079.4 for DOI-only documents. The least number of documents received by an organization is 569, and the highest number is 52,779, as opposed to 478 for the minimum number of documents with DOI and 41,997 for the maximum number of documents with DOI. The total number of citations obtained by 100 organizations with DOI is 68,66,250, which is 73.5% of total citations. An organization's minimum citations received is 849, its maximum citations is 8,76,753, and its average citations is 93,370.6. When just DOI-based citations are considered, the smallest citations are 636, the maximum citations are 6,56,860, and the average citations are 68,662.5. 13.7% of the citations are self-citations, with an average of 9372.9 per organization. There are 16 organizations (16%) that have received fewer than 1000 self citations, 39 organizations (39%) that have earned fewer than 5000 self citations, and 45 organizations (45%) that have received more than 10,000 self citations. The minimum number of self-citations is 100, while the greatest number is 85,490. The average  $h$ -index, which includes self-citations and all papers (both with and without DOI), is 81.5, with a minimum of 12 and a maximum of 246. If we only evaluate publications having a DOI and omit self-citations, the minimum  $h$ -index is 8, the maximum  $h$ -index is 203, and the average  $h$ -index is 66.5. This indicates that the  $h$ -index has decreased by 4 points at the minimum, 43 points at the maximum, and 15 points on average. According to the study, 30 organizations (30%) have a  $h$ -index difference point of less than 10, 69 organizations (69%) have a  $h$ -index difference point of less than 50, and one organization (1%) has a  $h$ -index difference point of greater than 50.

Table 4.2 shows the results of an additional examination of 100 organisations based on four types: *Universities*, *IITs*, *IEST*, *IISC & IISER*, and *NITs*. *IEST*,

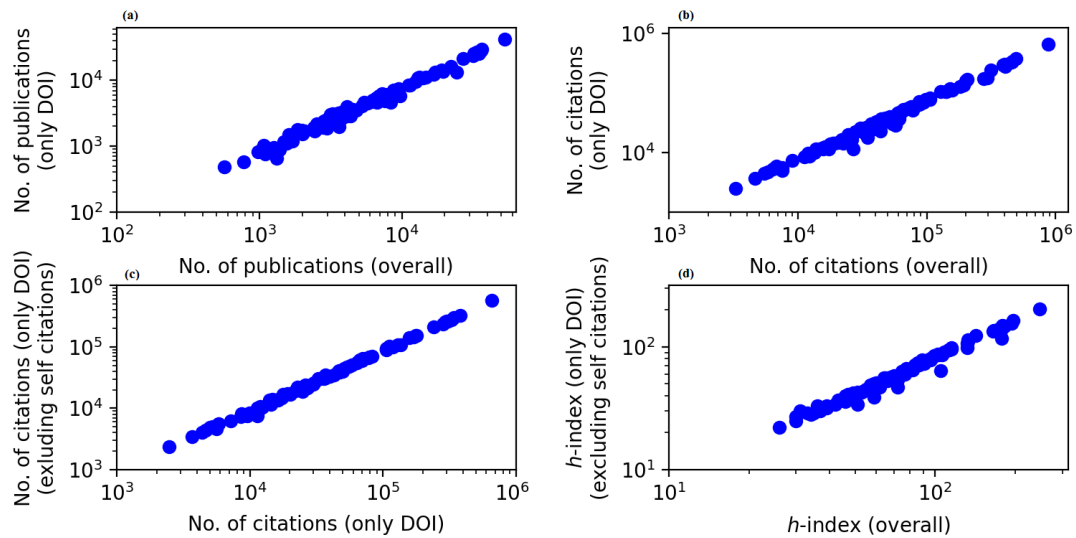


FIGURE 4.2: *Comparative analysis of 100 organizations based on DOI information.*

*IISC*, & *IISER* is on top with 81.7% of DOI documents, followed by *IITs* with 79.9% and *NITs* with 72.7%. *NITs* earn 76.6% of citations with valid DOIs, followed by *IITs* (74.9%), and *Universities* (72.6%) at the bottom. Self citations account for 13.9% of citations received by *IITs*, 13.7% by *Universities*, and 12.0% by *NITs*. *IITs* has an average *h*-index of 108.9, *IEST*, *IISC*, and *IISER* has an average *h*-index of 91.6, and *NITs* has an average *h*-index of 64.4. If we only evaluate DOI-based publications and omit self-citations, the average *h*-index of *IITs* drops to 20.6, *IEST*, *IISC*, and *IISER* drops to 16.9, and *NITs* drops to 10.0.

#### 4.1.5.3 Journal level bibliometrics

The comparison of (a) documents (with and without DOI), (b) citations (with and without DOI), (c) self citations and citations (with DOI), and (d) *h*-index is shown in Fig. 4.3 (with and without DOI and self citations). A total of 1000 journals are analyzed. We reviewed 14,15,093 documents and discovered that 11,87,692 of them have DOIs, accounting for 83.9% of the total. Only DOI-based documents reduced the number of documents in 77.6% of journals, such as 45.2% of journals with a difference of fewer than 100 documents, 32.4% of journals with a difference of greater than or equal to 100 documents, and so on. The total number of citations obtained

TABLE 4.2: *Comparative analysis of 100 organizations based on DOI information.*

Organization Types	No. of pubs	(%) of pubs (only DOIs)	No. of cites	(%) of cites (only DOIs)	(%) of self cites (only DOIs)	Avg. $h$ index	Avg. $h$ index (only DOIs, exc. self cites)
Universities	451489	73.8	4917831	72.6	13.7	74.5	62.0
IITs	236547	79.9	2993534	74.9	13.9	108.9	88.3
IEST, IISC & IISER	70063	81.7	1107018	73.4	13.2	91.6	74.7
NITs	39059	72.7	318676	76.0	12.0	64.4	54.4

by 1000 journals is 2,25,70,461, with 1,40,05,489 DOI citations accounting for 62.1% of all citations. 99.9% of journals experienced a decrease in citations, with 36.7% experiencing a decrease of less than 1000 citations, 29.1% experiencing a decrease of less than 5,000 citations, 13% experiencing a decrease of less than 10,000 citations, and 21.1% experiencing a decrease of more than or equal to 10,000 citations. 95.2% of journals have received self citations, with 54.2% having less than 500 self citations, 14.4% having less than 1000 self citations, 23.8% having less than 5000 self citations, and 7.6% having more than 5000 self citations. The average  $h$ -index, which includes self-citations and all papers (both with and without DOI), is 43.9, with a minimum of 2 and a maximum of 344. If we just evaluate DOI-based papers and omit self-citations, the average  $h$ -index is 31.6, with a minimum of 1 and a maximum of 236. Similarly, 52.6% of journals experienced a reduction of fewer than 10 points in their  $h$ -index, 28.5% experienced a decrease of less than 20 points in their  $h$ -index, and 18.9% experienced a difference of more than 20 points in their  $h$ -index.

Table 4.3 shows the results of a second examination of 1000 journals based on five disciplines: *Engineering*, *Social Sciences*, *Life Sciences*, *Sciences*, and *Humanities*. According to a preliminary analysis, *Engineering*, *Sciences*, *Humanities*, and *Social Sciences* disciplines have more than 80% of documents with DOIs, while *Life Sciences*

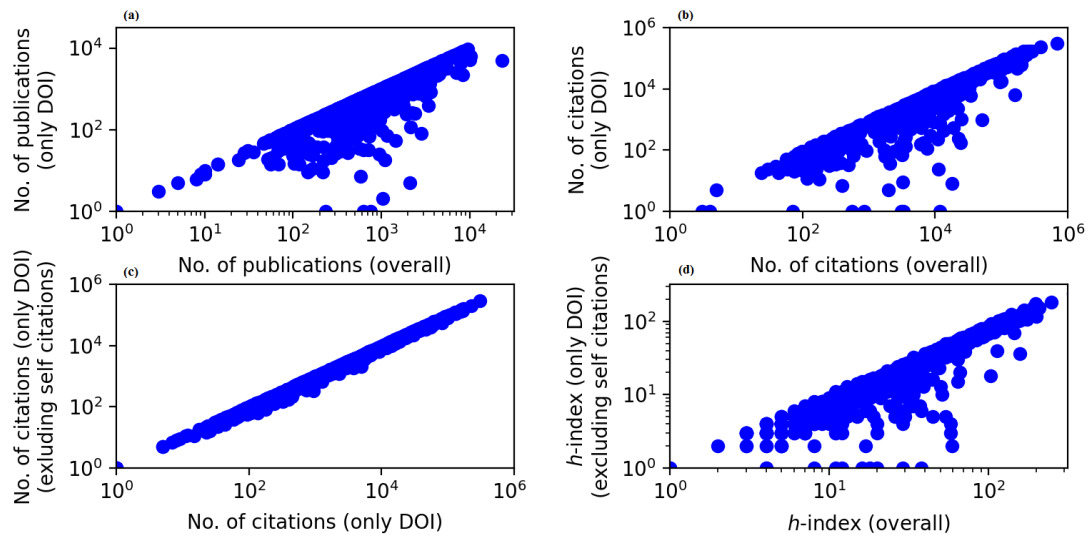


FIGURE 4.3: *Comparative analysis of 1000 journals based on DOI information.*

has only 60.6%. The *Engineering* discipline has the most citations of all, whereas the *Sciences* discipline has the most DOI citations (73.4%). The *Engineering* discipline, on the other hand, has the fewest DOI citations, with only 60.5%. The closest difference between the disciplines is in *Life Sciences*, where 60.6% of documents have DOIs and 67.0% of citations have DOIs, which is the closest difference. With 12.6% self-citations, *Sciences* takes the lead, followed by *Engineering* and *Social Sciences*. In comparison, *Engineering* received the most citations, while *Sciences* received the most self-citations. *Life Sciences* has the greatest average *h*-index, which includes self-citations and all documents (with and without DOIs), whereas *Social Sciences* has the lowest. *Engineering* has the highest average *h*-index drop of 13.2, while *Social Sciences* has the lowest average *h*-index decline of 7.6. The average drop in the *h*-index is 10 points across all disciplines.

## 4.2 Methodology

Here we have presented the methodology used to link indexing databases and the generation of unified informetrics for authors, organizations and journals.

TABLE 4.3: *Comparative analysis of 1000 journals based on DOI information.*

Journal Disciplines	No. of pubs	(%) of pubs (only DOIs)	No. of cites	(%) of cites (only DOIs)	(%) of self cites (only DOIs)	Avg. $h$ index	Avg. $h$ index (only DOIs, exc. self cites)
Engineering	1179771	85.7	19176940	60.5	10.5	45.5	32.3
Social Sciences	86719	80.4	1268836	70.7	9.5	29.6	21.9
Life Sciences	68142	60.6	858220	67.0	7.1	49.9	39.1
Sciences	53458	81.0	783108	73.4	12.6	48.8	38.6
Humanities	27003	84.3	483357	72.0	9.0	46.3	37.5

#### 4.2.1 Generation of doi based citation database

In this algorithm, an entity specification, i.e., an author, organization, or journal, will be required to input its credentials. For example, an author will input his Orcid ID, an organization can input its name, and a journal can input its ISSN. After receiving the valid input from the mentioned entities, the first step of extracting article information from multiple databases will be executed. The output of this step will provide resultant information fetched from multiple indexing databases on the basis of filtration. Filtration will be applied to the fetched data to retain the articles with DOIs and articles that do not have DOI information associated with them will be neglected. Thus, a merged article database will be created after the filtration step. This merged article database will be further queried by indexing databases like Scopus and Web of Science to extract the citation details of final articles. Two new databases will be created, one for Scopus and one for Web of Science. Citation data will be filtered again on the basis of DOIs and, thus, final results will be merged into a single database, called the merged citation database. Algorithm 2 describes the steps for extracting article and citation details of the entered entity for performing citation analysis.

---

**Algorithm 2** Generation of doi based citation database
 

---

**Require:** Entity identifier**Ensure:** doi based citation database

```

1: for each entity do
2:    $[A_i] \in DB_i$ , where  $i = 1, \dots, N, N > 0$ 
3:    $\triangleright$  /*  $[A_i]$  is list of articles, doi numbers in database  $DB_i$ . */
4:   for each doi in  $[A_i]$  do
5:      $[C_N]$ =list of citations
6:     for each citation in  $[C_N]$  do
7:       if doi exists then
8:          $[CD_i]=doi$ 
9:          $\triangleright$  /*  $[CD_i]$  is doi based citation database, computed on  $[A_i]$ . */
10:      end if
11:    end for
12:  end for
13: end for
14: Repeat step 1 and 2 to get  $A_1, \dots, A_N$  and  $CD_1, \dots, CD_N$  from  $DB_1, \dots, DB_N$ 
15:    $\triangleright$  /* merge all citation databases for a given entity. */
16:  $CD_{all} = CD_1 \cup CD_2 \cup \dots \cup CD_N$ 
17:  $\triangleright$  /* where  $CD_{all}$  contains only those citations for a given entity whose doi exists
    (including duplicates). */

```

---

### 4.2.2 Computation of weighted unified informetrics

The combined citation database generated in algorithm 2 will be utilized to perform citation analysis and the generation of Conflate informetric ledger. In the first step of citation analysis, common and unique citations among both indexing databases will be extracted. Unique citations from both databases are merged with common citations in both indexing databases to produce the final citation count of an entity, i.e., author, organization and journal. This final citation count is used to calculate the unified informetrics of an entity. For the consideration of these citations in Conflate, a concept of weighted informetrics is introduced, where common and unique citations will be assigned a weight. Algorithm 3 describes the process in sequence and Fig. 4.4 summarizes the computation of weighted unified informetrics for different entities.



---

**Algorithm 3** Computation of weighted unified informetrics
 

---

**Require:**  $CD_{all}$ : Conflate citation database

**Ensure:** Weighted unified informetrics and research indicators

```

1:  $CD_{common} = CD_1 \cap CD_2 \cap \dots \cap CD_{all}$ 
2:  $CD_{uniue} = CD_{all} - CD_{common}$ 
3: for each  $doi$  in  $[CD_{unique}]$  do
4:    $P = \text{Count}(doi)$  in  $CD_{all}$ 
5:    $W_{doi} = P/N$ 
6:                                     ▷ /* N is number of citation databases,  $N > 0$ . */
7: end for
8: Compute  $h$ -index
9: Display number of publications, citations and  $h$ -index for a given entity

```

---

### 4.2.3 The weighted unified informetrics algorithm

**The Weighted Unified Informetrics (WUI) Algorithm:** Bibliographic databases like Scopus and WoS are employed in the suggested technique because of their indexing age, data availability, and validity. The “Conflate” weighted unified informetrics system has been discussed and proposed (see Fig. 4.5).

## 4.3 Data description and filtering

Scopus and Web of Science are indexing databases that are being used worldwide. This also makes sense that they are considered a verified data source by government organizations, private sectors, ranking agencies, and academic institutions. These stakeholders also rely on the data provided by both platforms. To analyze the depth of the identified problem, data from both indexing databases has been fetched and analyzed.

### 4.3.1 Data sources

At the author level, profiles of different authors have been searched online. Different university websites are accessed and it has been found that ‘Monash University’, a public university in Melbourne, Australia, has provided its staff profiles at (<https://research.monash.edu/en/persons/>). A total of 6316 profiles exist with the names, designations, departments, and research contribution details of staff

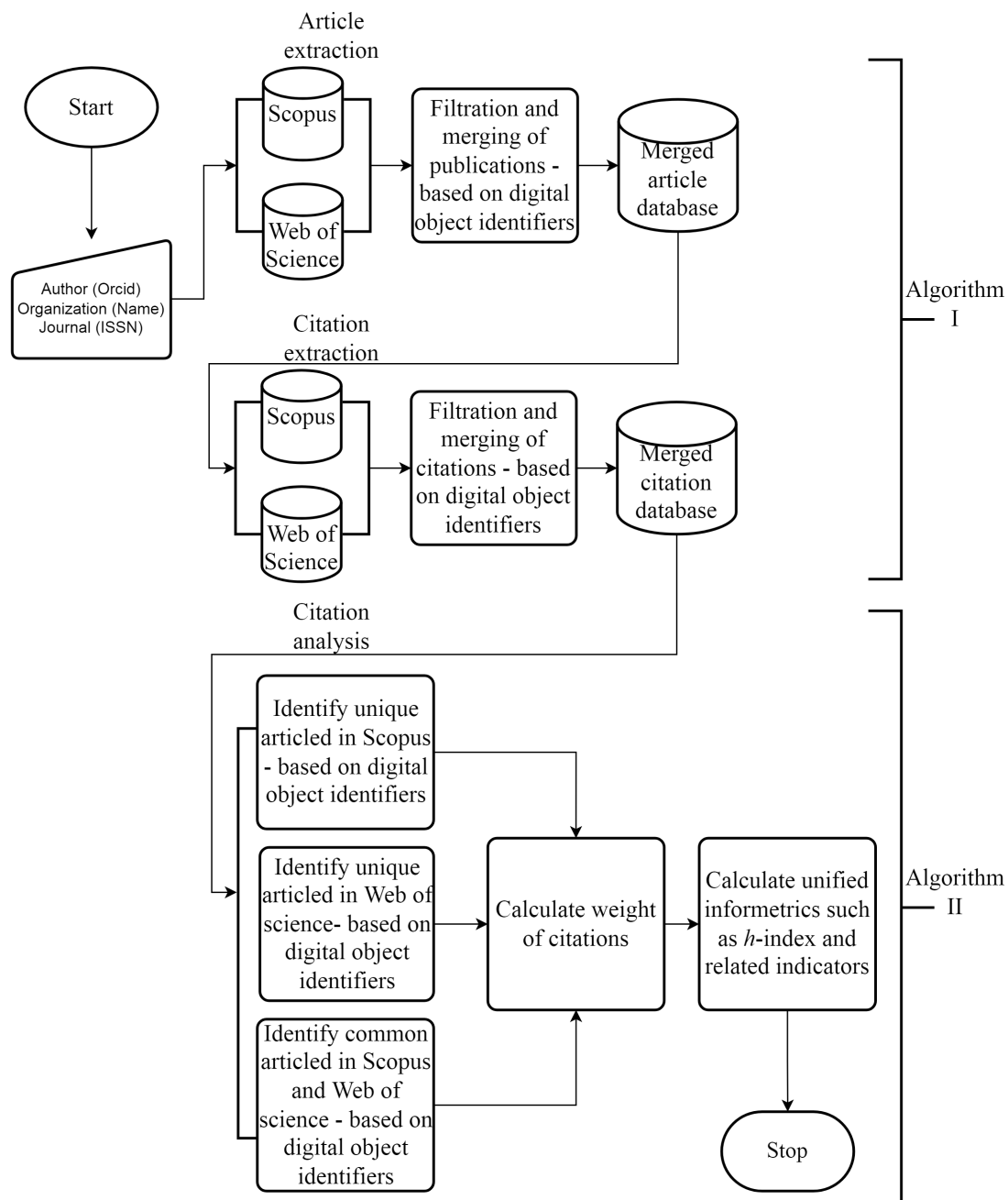


FIGURE 4.4: *Flowchart demonstrates the computation of weighted unified informetrics*

members and research scholars. Available profiles are searched, filtered and 400 profiles with the required information of author Orcid ID, Scopus ID, and Web of Science ID are selected. Selected profiles are identified from various disciplines, including medical sciences, engineering, agriculture, social sciences, humanities, etc. The approach is to identify the problem across multiple disciplines so that the intent of the problem

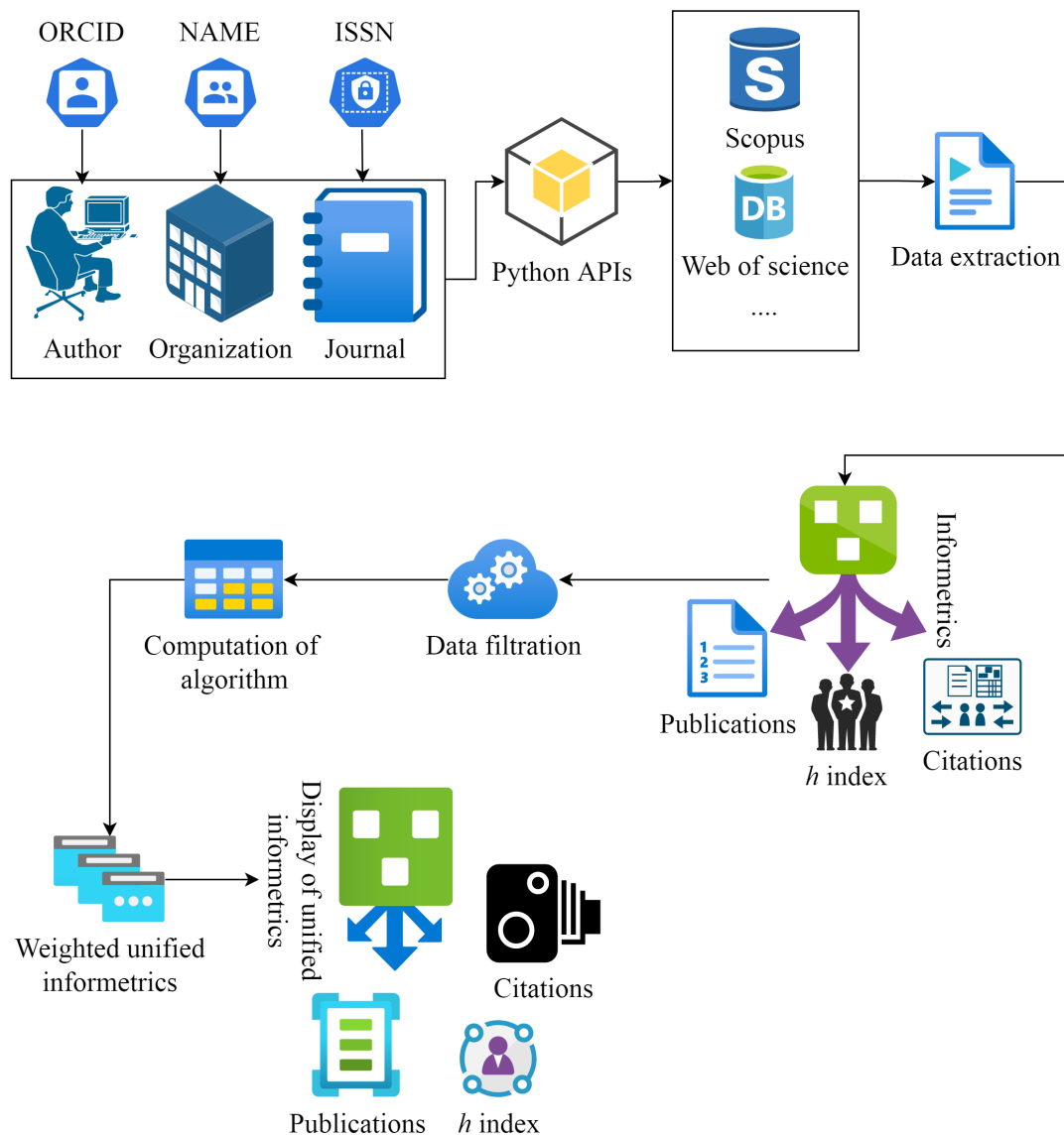


FIGURE 4.5: *Schematic representation of the proposed weighted unified informetrics*

may be observed deeply.

At the organization level, the website of NIRF (National Institutional Ranking Framework) (<https://www.nirfindia.org/2020/Ranking2020.html>) has been accessed. It shows rankings in different categories like, overall, university, engineering, management, pharmacy, college, medical, law, architecture, and dental. We have considered the overall category to cover almost all types of institutions. In the overall category, 200 institutions are listed, and we filtered and used the top 100 institutions in the overall category.

At the journal level, Scopus (<https://www.scopus.com/sources.uri>) and Web of Science (<https://jcr.clarivate.com/>) API's are used to fetch data. We observed different categories of sources, like journals, book series, and conference proceedings. The categories of journals are selected and their various disciplines like computer science, arts and humanities, physical sciences, health sciences, social sciences, and life sciences are observed. A random sample of 1195 journals, with major contributions from the field of computer science, has been selected. Filtration has been applied, and a sample of 1000 journals is considered for the analysis.

Fig. 4.6 shows the complete process of visiting the author's, organization's, and journal's profiles.

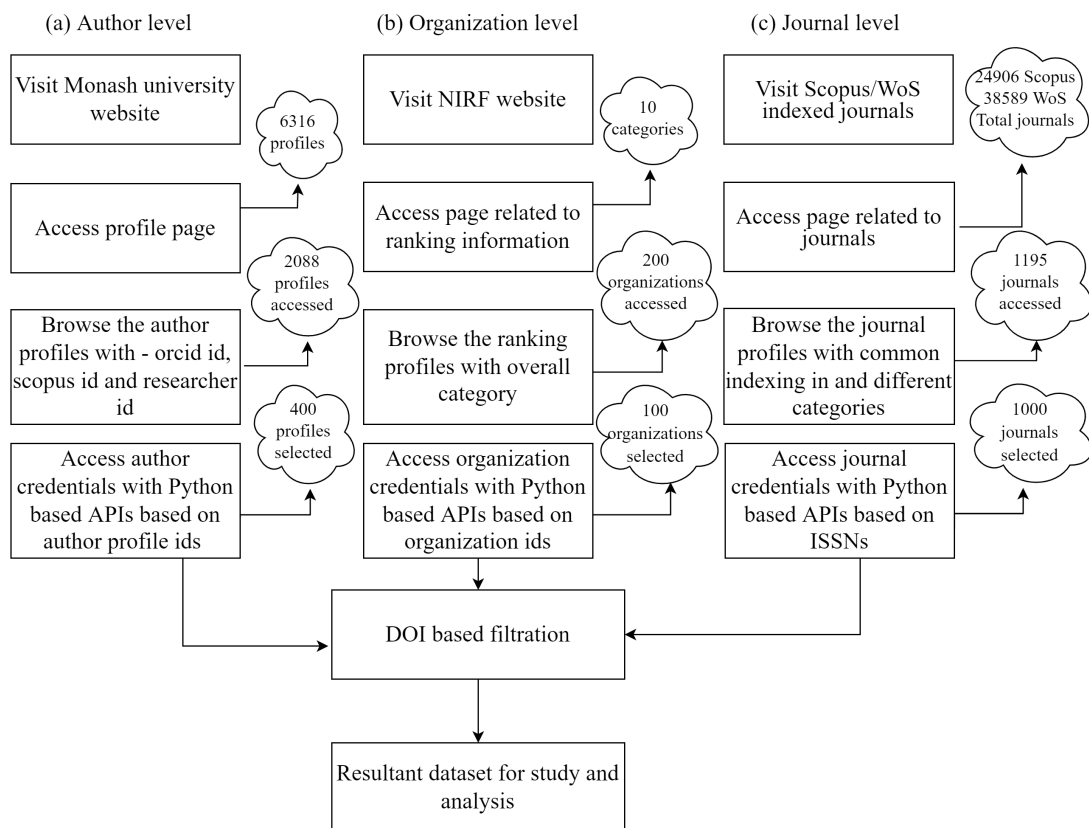


FIGURE 4.6: *Flowchart demonstrates the process of visiting the author's, organization's, and journal's profile.*

### 4.3.2 Data analytic

Indexing databases like Scopus and Web of Science were used for the work. Data from both indexing databases was fetched with the help of Python based APIs [130–132]. For identity specification, article extraction and citation extraction data were fetched on a real-time basis. For all three entities (author, organization, and journal), identifiers like Orcid, organization name, and ISSN were used. All entities were required to give these identifiers as an input to the system. After receiving the inputs from the entities, values were passed to the indexing databases, and article and citation information were fetched. As real-time data was used in the work, results were always complete, real insights were available, processes were agile, and outcomes were generated without any barriers. The interface powered by Ganache with the integration of the Truffle framework has been used to provide the application functionality to the work (<https://trufflesuite.com/ganache/>). Fig. 4.7 gives the insight details of concepts used for the complete work.

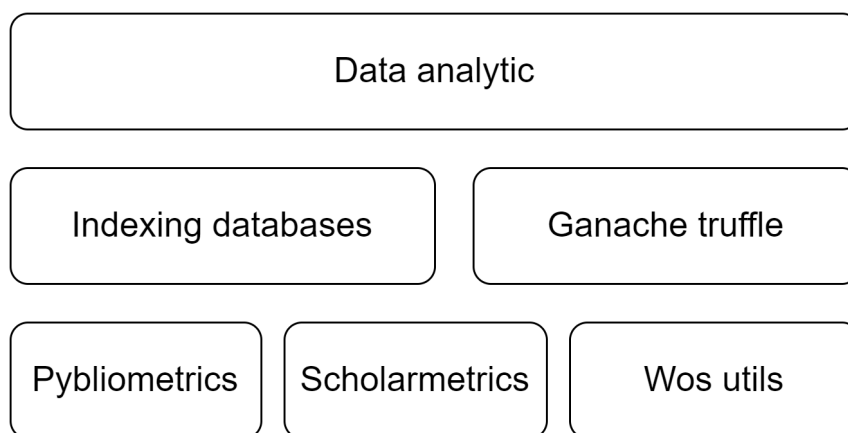


FIGURE 4.7: *Representation of concepts used to retrieve, compile, analyze and present the unified informetric ledger - Conflate.*

### 4.3.3 Article extraction and filtration

Here we have presented the detailed description of the article extraction process at author, organization and journal level.

### 4.3.3.1 Author level bibliometrics

An author will be required to input his credentials in the form of an Orcid ID. After receiving the valid input from an author, the system will connect to the first indexing database, i.e., Scopus, and fetch author details like author name, author id, affiliation name, affiliation city, and affiliation country etc. The system will also fetch his publication details like publication id, name, type, DOI, ISSN, volume, article number, page range, author keywords, citation count, funding accreditation, and funding number etc. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the Orcid ID of an author in the folder named as Rough Files/Authors/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, the system will connect to indexing database-2, i.e., Web of Science, and start fetching the required author credentials like author name, author ids, number of authors, publication details like publication name, publication id, publication type, DOI, journal name, publisher name, publisher address, and citation count etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the Orcid ID of an author in the folder named as Rough Files/Authors/Indexing database - 2.

In the next step, the system will process the results saved in the two csv files for both indexing databases and filtration will be applied. Filtration will be done on the basis of articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. DOIs are assigned to 91.5% of Scopus documents and 82.3% of WoS documents. Thus, a merged article database will be created with a new structure and results will be saved in a folder named as Merged Files/Authors/ORCID ID-1 of an author.

### 4.3.3.2 Organization level bibliometrics

An organization will be required to input its credentials in the form of its name. After receiving the valid input from an organization, system will connect to the first indexing database, i.e., Scopus, and fetch the affiliation details like id, name, city, and country of an organization. An organization can view all the records returned by the system and can select any record matching its credentials with a valid id. The system will again connect to the indexing database Scopus and fetch the records associated with the selected affiliation ID by an organization. The system will fetch the details like organization name, organization id, DOI, publication id, number of publications, and number of citations received by all the publications. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the ORG ID of an organization in the folder named as Rough Files/ORG/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, system will connect to indexing database -2 i.e. Web of Science, and start fetching the required organization credentials on the basis of organization name, like number of publications, publication id, DOI, and citation count etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the ORG ID of an organization in the folder named as Rough Files/ORG/Indexing database - 2.

In the next step, the system will process the results saved in the 2 csv files for both indexing databases and filtration will be applied. Filtration will be done on the basis of articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. DOIs are assigned to 83.6% of Scopus documents and 77.2% of WoS documents. Thus, a merged article database will be created with a new structure and the results will be saved in a folder named as Merged Files/ORG ID-1 of an organization.

### 4.3.3.3 Journal level bibliometrics

A journal will be required to input its credentials in the form of an ISSN. After receiving the valid input from a journal, the system will connect to the first indexing database, i.e., Scopus, and fetch the publication details like publication id, DOI, year of publication, number of publications, and number of citations received by all the publications. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the ISSN of a journal in the folder named as Rough Files/Journals/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, system will connect to indexing database-2, i.e., Web of Science, and start fetching the required journal credentials on the basis of ISSN like number of publications, publication id, DOI, publication date, and citation count etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the ISSN of an organization in the folder named as Rough Files/Journals/Indexing database - 2.

In the next step, the system will process the results saved in the 2 csv files for both indexing databases and filtration will be applied. Filtration will be done on the basis of articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. DOIs are assigned to 92.5% of Scopus documents and 84.2% of WoS documents. Thus, a merged article database will be created with a new structure and results will be saved in a folder named as Merged Files/Journals/ISSN-1 of a journal.

### 4.3.4 Citation extraction and filtration

Here we have presented a detailed description of the citation extraction process at the author, organization and journal level.



#### 4.3.4.1 Author level bibliometrics

During article extraction and filtration, an author has passed his Orcid ID as an input to the system. On the basis of the input, article details were fetched from multiple databases and results were stored in a merged article database. Now the merged article database will be used as an input for citation extraction and filtration. It contains a column named "source" which specifies the source of data in the file. In our case, the source can be Scopus or it can be the Web of Science. The system will connect to the indexing database, i.e., Scopus, and extract the required citation details for each publication of an author on the basis of DOI. The system will extract citation information like publication id and DOI etc. For example, if an author has 5 publications with 2 citations each, then 10 records will be extracted by the query and saved publication-wise. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the Orcid ID of an author in the folder named as Cite Files/Authors/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, system will connect to indexing database-2, i.e., Web of Science, and start fetching the required citation details like their publication id and DOI etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the Orcid ID of an author in the folder named as Cite Files/Authors/Indexing database - 2.

In the next step, the system will process the results saved in the two csv files for both indexing databases and filtration will be applied. Filtration will be done on the basis of cited articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. Thus, a merged citation database will be created with a new structure and results will be saved in a folder named as Merged Files/Authors/ORCID ID-2 of an author.

#### 4.3.4.2 Organization level bibliometrics

During article extraction and filtration, an organization has passed its name as an input to the system. On the basis of the input, article details were fetched from multiple databases and results were stored in a merged article database. Now the merged article database will be used as an input for citation extraction and filtration. It contains a column named "source" which specifies the source of data in the file. In our case, the source can be Scopus or it can be the Web of Science. The system will connect to the indexing database, Scopus, and extract the necessary citation details for each publication of an organisation based on the DOI. The system will extract citation information like publication id and DOI etc. For example, if an organization has 50 publications with 5 citations each, then 250 records will be extracted by the query and will be saved publication wise. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the ORG ID of an organization in the folder named as Cite Files/ORG/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, the system will connect to indexing database -2 i.e. Web of Science, and start fetching the required citation details like their publication id and DOI etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the ORG ID of an organization in the folder named as Cite Files/ORG/Indexing database - 2.

The system will then process the results saved in the two csv files for both indexing databases and apply filtration. Filtration will be done on the basis of cited articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. Thus, a merged citation database will be created with a new structure and results will be saved in a folder named as Merged Files/ORG/ORG ID-2 of an organization.

#### 4.3.4.3 Journal level bibliometrics

During article extraction and filtration, a journal has passed its ISSN as an input to the system. On the basis of the input, article details were fetched from multiple databases and results were stored in a merged article database. Now the merged article database will be used as an input for citation extraction and filtration. It contains a column named "source" which specifies the source of data in the file. In our case, the source can be Scopus or it can be the Web of Science. The system will connect to the indexing database, i.e., Scopus, and extract the required citation details for each publication of a journal on the basis of DOI. The system will extract citation information like publication id and DOI etc. For example, if a journal has 200 publications with 5 citations each, then 1000 records will be extracted by the query and will be saved publication-wise. After fetching the required data from the first indexing database, a csv file will be created and data will be saved with the ISSN of a journal in the folder named as Cite Files/Journals/Indexing database - 1.

After fetching the complete details from indexing database-1, i.e., Scopus, the system will connect to indexing database-2, i.e., Web of Science, and start fetching the required citation details like their publication id and DOI etc. After fetching the required data from the second indexing database, a csv file will be created and data will be saved with the ISSN of a journal in the folder named as Cite Files/Journals/Indexing database - 2.

In the next step, the system will process the results saved in the two csv files for both indexing databases and filtration will be applied. Filtration will be done on the basis of cited articles with DOI numbers only. Primarily, articles with DOI numbers will be retained, and articles that do not have any DOI information associated with them will be neglected. Thus, a merged citation database will be created with a new structure and results will be saved in a folder named as Merged Files/Journals/ISSN-2 of a journal.

## 4.4 Citation analysis and unified informetrics

Here we have presented a detailed description of the citation analysis process and the calculation of unified infometrics at the author, organization and journal level.

### 4.4.1 Author level bibliometrics

During citation extraction and filtration, citations for an entity author were fetched from indexing database -1, i.e., Scopus, and indexing database -2, i.e., Web of Science. These fetched citations were analyzed and filtration was applied to the results. After applying filtration, a merged citation database was created and it was saved with a structure containing the source, i.e., whether the citation is fetched from Scopus or from the Web of Science, the Orcid ID of an author, the DOI of the main publication, the publication id of the main publication, and the DOI of citations. This compiled data was saved in a folder named as Merged Files/Authors/ORCID ID-2 of an author.

In the next step, the complete data was divided into three parts on the basis of indexing database information stored in the Source column. In the first part, all publications available uniquely in Scopus were extracted. In the second part, all publications available uniquely on the Web of Science were extracted. In the third part, all publications that were common in both indexing databases, i.e., Scopus and Web of Science, were extracted. All three parts were fetched, and payoff weight was calculated for all publications one by one. After applying payoff weight to the number of citations, the final citation count for an author was calculated. In the last step, final publications with final citation count of each publication were saved in the folder named as Result Files/Mine/Authors/ORCID ID of an author, and DOIs of the final citation count were saved in Result Files/Cites/Authors/ORCID ID of an author.

This final publication and citation count were further used to calculate unified informetrics. As an output to the author, the number of publications, total number of citations,  $h$ -index of an author, self-citations of an author, repeated citations of an author, and actual citations of an author were displayed. The input for this generated

output was just an Orcid ID of an author as a step -1 of the system. The system accessed, processed, and analyzed unified informetrics in real time, and results were produced. Authors can see their single publication, single citation, and single  $h$ -index as an output across multiple indexing databases.

#### 4.4.2 Organization level bibliometrics

During citation extraction and filtration, citations for an entity organization were fetched from indexing database-1, i.e., Scopus, and indexing database-2, i.e., Web of Science. These fetched citations were analyzed and filtration was applied to the results. After applying filtration, a merged citation database was created and it was saved with a structure containing the following: source, i.e., whether the citation is fetched from Scopus or from the Web of Science; organization name and ID; DOI of main publication, publication id of main publication; and DOI of citations. This compiled data was saved in a folder named as Merged Files/Org/ORG ID-2 of an organization.

In the next step, the complete data was divided into three parts on the basis of indexing database information stored in the source column. In the first part, all publications available uniquely in Scopus were extracted. In the second part, all publications available uniquely on the Web of Science were extracted. In the third part, all publications that were common in both indexing databases, i.e., Scopus and Web of Science, were extracted. All three parts were fetched and pay off weight was calculated for all publications one by one. After applying pay off weight to number of citations, final citation count for an organization was calculated. In the last step, final publications with final citation count of each publication were saved in the folder named as Result Files/Mine/ORG/ORG ID of an organization and DOIs of the final citation count were saved in Result Files/Cites/ORG/ORG ID of an organization.

This final publication and citation count were further used to calculate unified informetrics. As an output to the organization, the number of publications, total number of citations,  $h$ -index of an organization, self-citations of an organization, repeated

citations of an organization, and actual citations of an organization were displayed. Input for this generated output was just the name of an organization as a step -1 of the system. The system accessed, processed, and analyzed unified informetrics in real time, and results were produced. Organizations can see their single publication, single citation, and single  $h$ -index as an output across multiple indexing databases.

### 4.4.3 Journal level bibliometrics

During citation extraction and filtration, citations for an entity journal were fetched from indexing database-1, i.e., Scopus, and indexing database-2, i.e., Web of Science. These fetched citations were analyzed and filtration was applied to the results. After applying filtration, a merged citation database was created and it was saved with a structure containing Source, i.e., whether the citation is fetched from Scopus or from the Web of Science, ISSN, DOI of main publication, publication id of main publication, publication date, and DOI of citations. This compiled data was saved in a folder named as Merged Files/Journals/ISSN-2 of a journal.

In the next step, the complete data was divided into three parts on the basis of indexing database information stored in the source column. In the first part, all publications available uniquely in Scopus were extracted. In the second part, all publications available uniquely on the Web of Science were extracted. In the third part, all publications that were common in both indexing databases, i.e., Scopus and Web of Science, were extracted. All three parts were fetched, and payoff weight was calculated for all publications one by one. After applying payoff weight to the number of citations, the final citation count for a journal was calculated. In the last step, final publications with final citation count of each publication were saved in the folder named as Result Files/Mine/Journal/ISSN of a journal, and DOIs of the final citation count were saved in Result Files/Cites/Journal/ISSN of a journal.

This final publication and citation count were further used to calculate unified informetrics. The number of publications, total number of citations,  $h$ -index of a journal, self-citations of a journal, repeated citations of a journal, and actual citations

of a journal were displayed as an output to the journal. The input for this generated output was just the ISSN of a journal as a step -1 of the system. The system accessed, processed, and analyzed unified informetrics in real time, and results were produced. Journals can see their single publication, single citation, and single  $h$ -index as an output across multiple indexing databases.

The year of publication, final publication count, and final citation count were also used to calculate the impact factor of a journal. The system prompted the journal to enter the number of years for which impact was required. The default value was set to 2. As per input given by the journal, results were calculated and the impact factor was displayed with additional information like the number of previous years, i.e. 2 or more, entered year(for which the impact was required), number of publications, and number of citations.

## 4.5 Discussion and summary

At the beginning of this chapter, we talked about the different entities associated with citation analysis. The entities were categorized into three broad categories and the linkage analysis was started. Different indexing databases use different terminologies to keep track of the scientific work of an author, organization and journal. Hence, there was a requirement to provide uniqueness to all entities. For authors, we used Orcid as an identifier, for organization, we used organization ID; and for journals, we used ISSN.

The next step was to retrieve the data for the analysis. For the retrieval of data, there was a requirement to identify sources. We identified that different stakeholders, like government agencies, accreditation bodies, and ranking agencies, have enormous trust in Scopus and the Web of Science. So we considered these two as valid sources of information for the retrieval of the data. Data retrieval was done from Scopus and Web of Science on the basis of Python based APIs. Data retrieval was also initiated in three categories, authors, organizations and journals. The complete process of data retrieval was divided into two steps of extraction, i.e., article level and citation level.

After completing the data retrieval, the next step was to perform filtration of the required data from the complete database. Filtration was done on the basis of DOI. Filtration was applied to the complete database of all three entities, i.e., on author, organization and journal. The next step was to apply filtration to the citation data retrieved from indexing databases.

Citation analysis was done on the filtered data where citations from both indexing databases, i.e., Scopus and Web of Science, were fetched. Citation analysis requires rigorous calculation at different steps to provide a single publication and a citation count of all three entities, like author, organization and journal. This single publication and citation count were used to further calculate unified informetrics, which presented a single index to the scientific work.

While performing citation analysis and calculating unified informetrics, a concept of weighted unified informetrics was used. This concept provided a novel feature in the calculation of unified informetrics as it added a mechanism of giving weight to the citations at different levels, i.e. unique citations in indexing database -1, i.e., Scopus, unique citations in indexing database -2, i.e., Web of Science, and common citations in both indexing databases, i.e., Scopus and Web of Science.

In the next chapter, statistical analysis of Conflate (Unified Informetrics) generated for three entities: author, organization, and journal, is discussed.



---

---

## CHAPTER 5

---

# Statistical analysis of Conflate (unified informetrics (UI))

This chapter discusses unified informetrics generated for three entities: author, organization, and journal. The question of interest is how the result of generated unified informetrics is different from the traditional methods of citation analysis. Three databases, such as Scopus, Web of Science and Conflate, are used. Finally, the results are broken down into three categories: publications, citations, and the  $h$  index.

### 5.1 Author level bibliometrics

Conflate, a combination of two or more sets of information, has presented a novel approach to preserve the features of two indexing databases, i.e., Scopus and Web of Science. By combining the features of both indexing databases, Conflate has also represented itself as a single stated measure to calculate the impact of scientific work by an author, organization, and journal. The number of publications, number of citations, and  $h$ -index are presented in a three-tier architecture. The purpose of

considering these three parameters for the analysis is the fact that they are directly connected with the scientific work of an author, organization and journal.

Different stakeholders, like government agencies, ranking agencies, career assessment agencies, and accreditation agencies, are also keen to know about these parameters only. Scientific influence, impact, contribution, and collaboration in human and scientific societies also depend on these parameters. Individuals get recognition for their scientific work, organizations get recognition of their scientific influence, and journals get scientific impact in human society with these three parameters. For journals, four parameters are considered instead of the basic three. The fourth one is the impact factor. It is considered a very common way to check the influence of a journal in a scientific society.

Authors, organizations and journals also cite their own scientific contributions. This results in a different perspective of thinking to promote self-scientific work for its deserving recognition at the initial stages of its publication. This scenario can be observed very easily among different entities like authors, organizations, and journals. Analysis with self-citations is also presented in context with Scopus, Web of Science, and Conflate [133]. Finally, repetitions in citations are analyzed. For example, how many times have different entities like authors, organizations, and journals cited a particular publication in their scientific work [134]. There could be various perspectives behind it. Such citation repetition may be related to self-citations of these entities as well.

The primary identifier used to maintain uniqueness among all authors for author level analysis is Orcid ID (<https://orcid.org/>). It is a commonly used identifier to distinguish the authors from one another. Two indexing databases, i.e., Scopus and Web of Science, were used to retrieve the publication and citation details of authors. Scopus uses the concept of Scopus author ID and the Web of Science uses the concept of researcher ID to maintain the uniqueness of authors. But in Conflate, features and outcomes of both indexing databases are combined to identify common elements among both indexing databases.

The answer to this problem was found in Orcid ID. The website of Monash University was accessed to retrieve various profiles at different levels. For example, professor, associate professor, senior lecturer, or lecturer, etc. While retrieving the details of various authors, profiles were specifically checked for the availability of Orcid ID. In the next step, on the basis of Orcid ID, profiles were retrieved which carried both Scopus author ID and Web of Science researcher ID in the database.

In the final step, various details were retrieved from both indexing databases using the Scopus author ID and the Web of Science researcher ID, such as author discipline or subject area, publication count, citation count, and  $h$ -index. Initially, 6316 author profiles were there, but after completing the above listed filtration steps, 400 author profiles were finalized, which had the data from both indexing databases with discipline/subject area information of authors as well. Fig. 5.1 gives the overview of the filtration process of author profiles. Filtered author profiles (400) were categorized



FIGURE 5.1: *Filtration process listing all the steps, from random author profiles to final list of author profiles.*

on the basis of disciplines/subject areas of authors. Fig. 5.2 gives the overview of 400 authors on the basis of their disciplines/subject areas. Social Sciences (66), Sciences (43), Humanities (20), Life Sciences (211), and Engineering (60).

Life Sciences	Social Sciences	Engineering	Sciences	Humanities
• 211	• 66	• 60	• 43	• 20

FIGURE 5.2: *Discipline/Subject area details of 400 authors.*

### 5.1.1 Number of publications

Fig. 5.3 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of publications of 400 authors. Scopus has reported the highest publication count for social sciences, sciences, humanities, and engineering, whereas the Web of Science has reported the highest publication count for life sciences. Conflate reported a publication count in the Scopus and Web of Science range for all disciplines except life sciences. In Fig. 5.4 during the comparative analysis

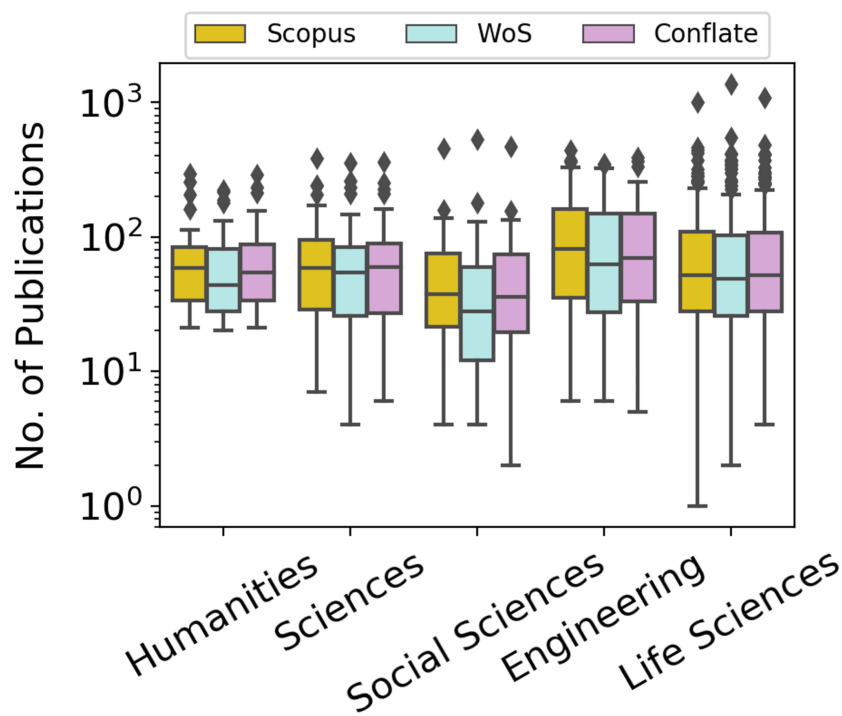


FIGURE 5.3: *A comparison of publications of 400 authors based on Scopus, Web of Science and Conflate.*

of the number of publications featured in Scopus, it is observed that the average number of publications published by an author is 83, whereas in Conflate it is 81. In the Web of Science, the average number of publications published is 79, as compared to an average of 81 publications per author in Conflate.

Table. 5.1 represents the comparative analysis of publications from Scopus, Web of Science, and Conflate for 400 author profiles among different disciplines.

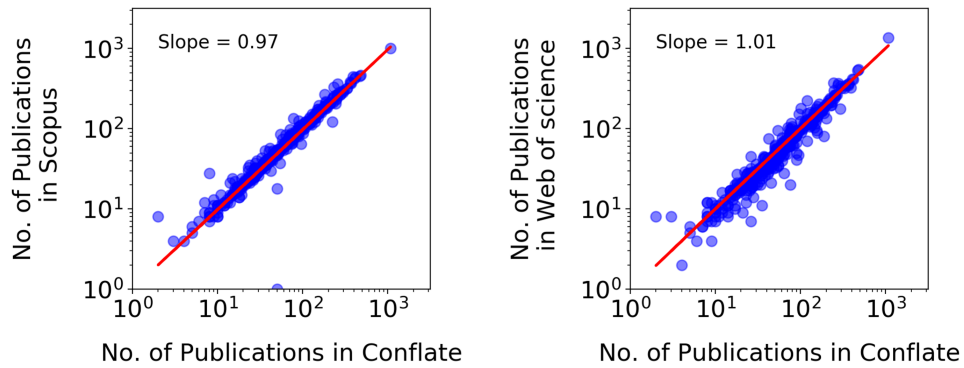


FIGURE 5.4: *Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level.*

Authors - 400 (Disciplines)	Publications		
	Scopus	Web of Science	Conflate
Life sciences	17793	18257	17951
Social sciences	3531	3113	3457
Engineering	6741	5658	5940
Sciences	3397	3187	3340
Humanities	1720	1517	1688

TABLE 5.1: *Comparative analysis of publications - author level*

### 5.1.2 Number of citations

Fig. 5.5 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of citations of 400 authors. The highest number of citations is reported in the discipline of life sciences in Scopus and the lowest number of citations is reported in the discipline of social sciences in Web of Science. For sciences, engineering, and life sciences, Conflate has reported the highest number of citations as compared to both Scopus and the Web of Science. The number of citations reported by Conflate for the remaining disciplines falls somewhere between Scopus and Web of Science. In Fig. 5.6 during the comparative analysis of the number of citations featured in Scopus, it is observed that the average number of citations received by an author is 2744, whereas in Conflate it is 2826. The average number of citations published on the Web of Science is 2562, as compared to 2826 in Conflate. The Web of Science has reported the lowest citations, whereas Conflate has reported the highest.

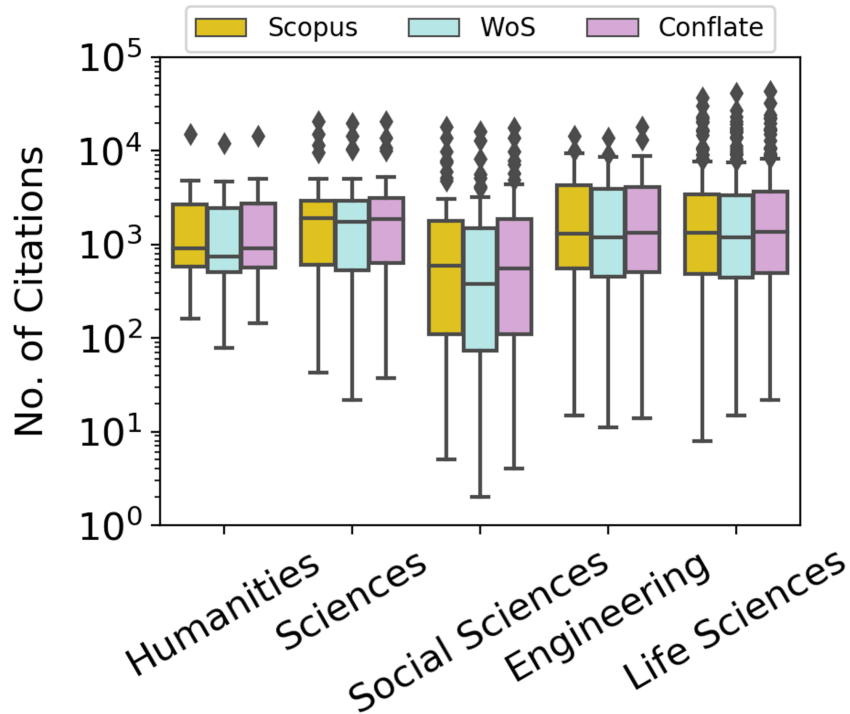


FIGURE 5.5: *A comparison of citations of 400 authors based on Scopus, Web of Science and Conflate.*

Table. 5.2 represents the comparative analysis of citations from Scopus, Web of Science, and Conflate for 400 author profiles among different disciplines.

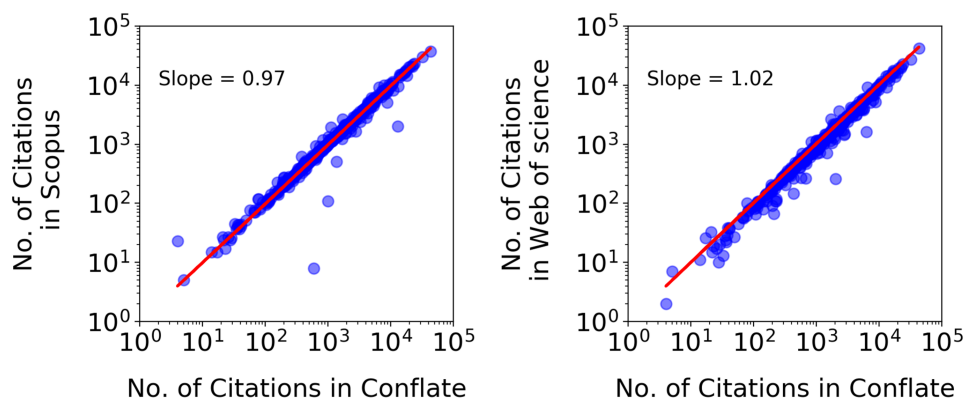


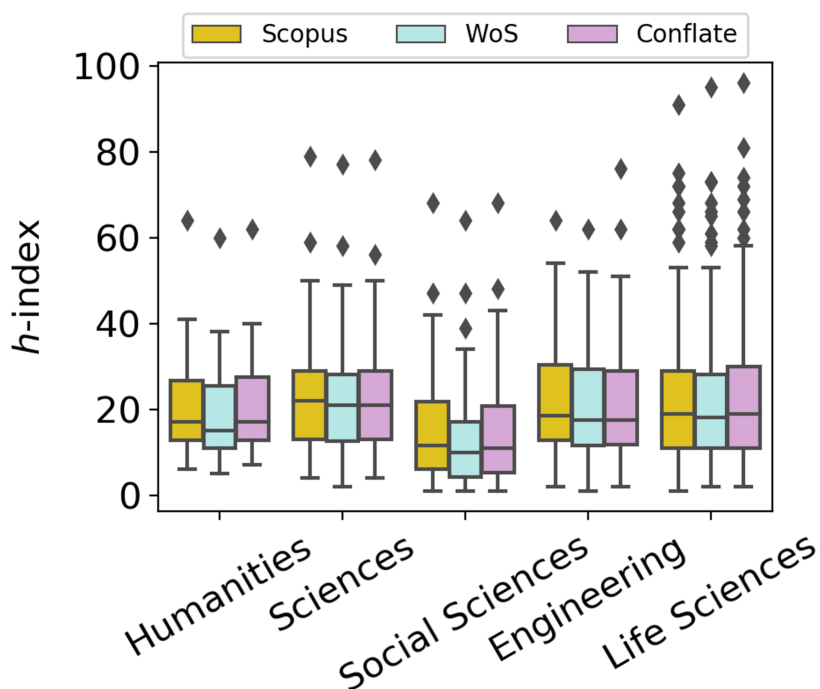
FIGURE 5.6: *Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level.*

Authors - 400 (Disciplines)	Citations		
	Scopus	Web of Science	Conflate
Life sciences	647698	631244	680537
Social sciences	115904	94195	114158
Engineering	161092	138631	162218
Sciences	127009	121752	128423
Humanities	45743	38986	44970

TABLE 5.2: *Comparative analysis of citations - author level*

### 5.1.3 Measuring the $h$ -index

Fig. 5.7 shows the comparison of results generated with Scopus, Web of Science and Conflate on the basis of the number of publications and citations of 400 authors. For the  $h$ -index of 400 authors, it was found that Conflate has reported the same  $h$ -index in social sciences and science discipline as reported by Scopus. For humanities and engineering, Conflate has reported  $h$ -index in the range of Scopus and Web of Science. For life sciences, Scopus and Web of Science have reported the same  $h$ -index whereas Conflate has reported one point higher than both. In Fig. 5.8 during the

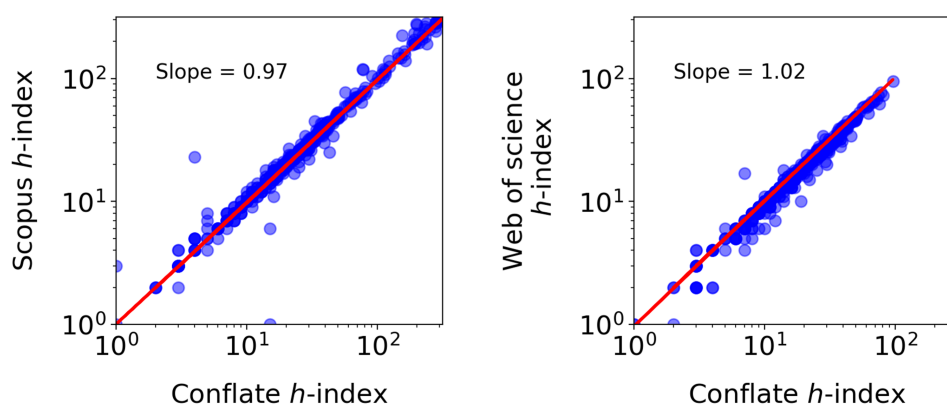
FIGURE 5.7: *A comparison of  $h$ -index of 400 authors based on Scopus, Web of Science and Conflate.*

Authors - 400 (Disciplines)	Average $h$ -index		
	Scopus	Web of Science	Conflate
Life sciences	22	22	23
Social sciences	15	13	15
Engineering	23	21	22
Sciences	24	23	24
Humanities	22	20	22

TABLE 5.3: *Comparative analysis of average  $h$ -index - author level*

comparative analysis of  $h$ -index featured in Scopus, it is observed that the average  $h$ -index received by an author is 21, whereas in Conflate it is 22. The average  $h$ -index received in the Web of Science is 20, as compared to 22 in Conflate. The Web of Science has reported the lowest average  $h$ -index whereas Conflate has reported the highest.

Table. 5.3 represents the comparative analysis of average  $h$ -index from Scopus, Web of Science, and Conflate for 400 author profiles among different disciplines.

FIGURE 5.8: *Comparative analysis based on  $h$ -index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at author level.*

#### 5.1.4 Self-citations vs. total-citations

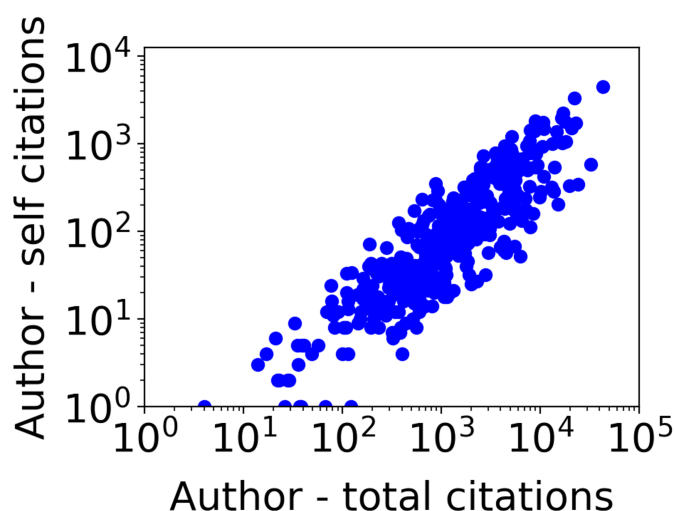
Comparative analysis of 400 authors on the basis of their total and self citations in Fig. 5.9 states that the average number of total citations for authors is 2825.76 and the average number of self citations for authors is 226.23. It can be concluded here that approximately 12.49 citations in the case of authors are self-citations, and it is



Authors - 400 (Disciplines)	Self citations	Total citations
Life sciences	49952	680537
Social sciences	6159	114158
Engineering	17423	162218
Sciences	12409	128423
Humanities	4548	44970

TABLE 5.4: *Comparative analysis of self citations vs. total citations - author level*

quite a high number of self-citations observed for authors. Table. 5.4 represents the

FIGURE 5.9: *A comparison of total and self citations of 400 authors based on Scopus, Web of Science and Conflate.*

comparative analysis of total citations and self-citations for 400 author profiles among different disciplines.

### 5.1.5 Repeated-citations vs. total-citations

A comparative analysis of 400 authors on the basis of their total and repeated citations in Fig. 5.10 states that the average number of total citations for authors is 2825.76 and the average number of repeated citations for authors is 655.10. It can be concluded here that approximately 4.31 citations in the case of authors are repeated citations. Table. 5.5 represents the comparative analysis of total citations and repeated citations for 400 author profiles across different disciplines.

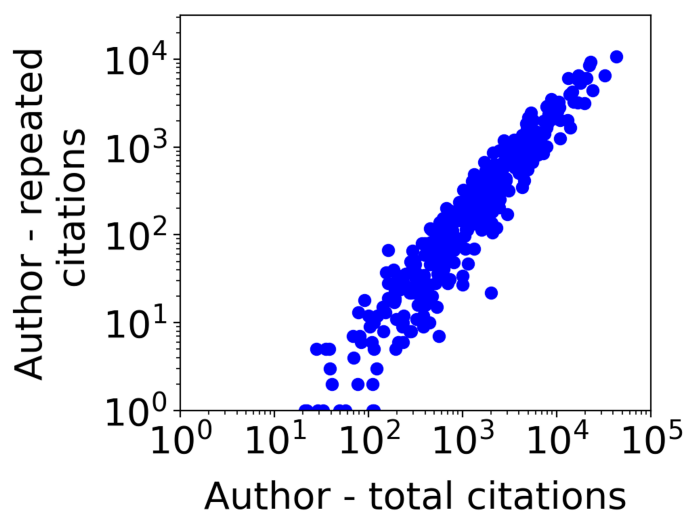


FIGURE 5.10: *A comparison of total and repeated citations of 400 authors based on Scopus, Web of Science and Conflate.*

Authors - 400 (Disciplines)	Repeated citations	Total citations
Life sciences	150334	680537
Social sciences	25495	114158
Engineering	44780	162218
Sciences	31287	128423
Humanities	10146	44970

TABLE 5.5: *Comparative analysis of repeated citations vs. total citations - author level*

### 5.1.6 Actual-citations vs. total-citations

Comparative analysis of 400 authors on the basis of their actual, self, repeated and total citations in Fig. 5.11 states that the average number of total citations for authors is 2825.76, average number of repeated citations for authors is 655.10; the average number of self-citations for authors is 226.23; and the average number of actual citations is 1944.43. Table. 5.6 represents the comparative analysis of actual citations, self-citations, repeated citations, and total citations for 400 author profiles among different disciplines.

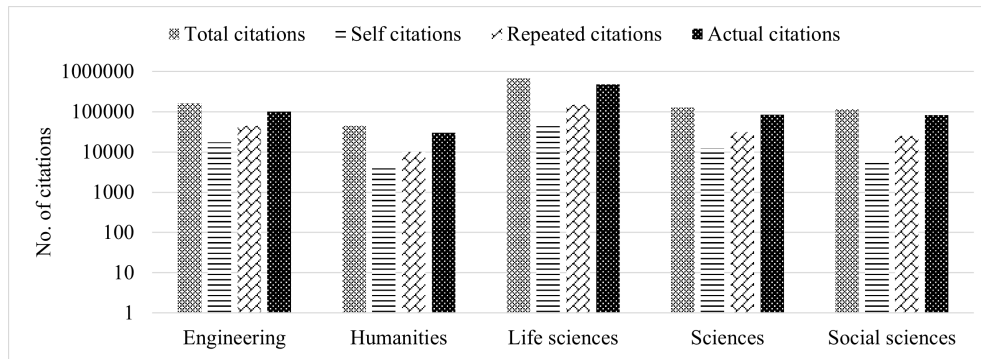


FIGURE 5.11: A comparison of total, self, repeated and actual citations of 400 authors based on Scopus, Web of Science and Conflate.

Authors - 400 (Disciplines)	Actual citations	Self citations	Repeated citations	Total citations
Life sciences	480251	49952	150334	680537
Social sciences	82504	6159	25495	114158
Engineering	100015	17423	44780	162218
Sciences	84727	12409	31287	128423
Humanities	30276	4548	10146	44970

TABLE 5.6: Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - author level

### 5.1.7 No. of citations vs. average $h$ -index

Fig. 5.12 shows the comparative analysis of citations with  $h$ -index for authors among multiple indexing databases [135]. During citation level analysis of authors in Scopus, it is observed that the average number of citations is 2743.61 and the average  $h$ -index earned is 21.29. It can be stated here that average cost of  $h$ -index is 128.87 per citation. When analysing the Web of Science results, it is discovered that the average number of citations is 2562.02, compared to the average  $h$ -index of 20.28, implying that the average cost is 126.33, which is less than the average cost of 128.87 calculated in Scopus. For Conflate, the average number of citations for authors is 2825.76, against the average  $h$ -index of 21.5, costing around 131.43 per  $h$ -index as compared to 126.33 in the Web of Science and 128.87 in Scopus. Table. 5.7 represents the comparative analysis of total citations, and average  $h$ -index for 400 author profiles among different disciplines.

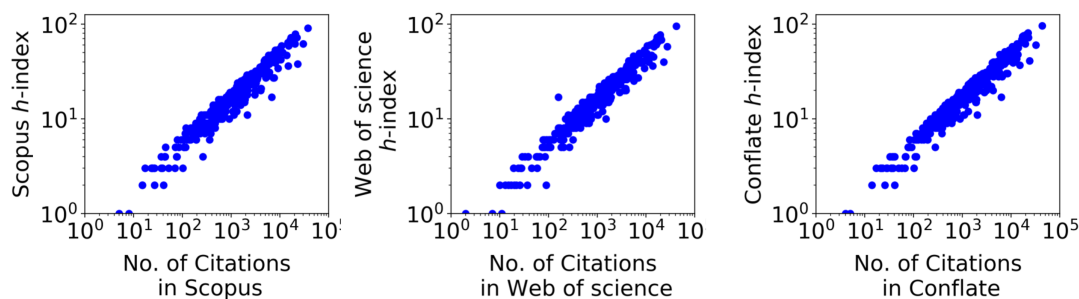


FIGURE 5.12: *A comparison of citations and h-index of 400 authors based on Scopus, Web of Science and Conflate.*

Indexing databases	Scopus		Web of Science		Conflate	
	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index
<b>Authors - 400 (Disciplines)</b>						
<b>Life sciences</b>	647698	22	631244	22	680537	23
<b>Social sciences</b>	115904	15	94195	13	114158	15
<b>Engineering</b>	161092	23	138631	21	162218	22
<b>Sciences</b>	127009	24	121752	23	128423	24
<b>Humanities</b>	45743	22	38986	20	44970	21

TABLE 5.7: *Comparative analysis of no. of citations vs. average h-index - author level*

## 5.2 Organization level bibliometrics

For organization level analysis, different platforms were accessed to fetch the list of organizations in India. The University Grants Commission (UGC)(<https://www.ugc.ac.in/>), Ministry of Human Resource Development(MHRD)(<https://www.education.gov.in/en>),and All India Council for Technical Education (AICTE) (<https://www.aicte-india.org/>) were the primary sources.

After exploring the data available on these platforms, a need was identified, to have a single source from where all kind of organizations may be covered which are continuously working towards higher education system in India. The National Institutional Ranking Framework (NIRF), an initiative of MHRD, was explored and a list of the top 100 institutions in India for the purpose of citation analysis was retrieved. These 100 organizations were further divided into 4 primary categories of

NITs, IITs, IEST, IISC and IISERs and Universities.

Two indexing databases i.e. Scopus, and Web of Science, were used to retrieve the publication and citation details of organizations. Scopus uses the concept of organization identifier and Web of Science uses the concept of organization name to maintain the uniqueness of organizations. But in Conflate, features and outcomes of both indexing databases were combined. The answer to the identified problem was found in the organization ID maintained by Scopus. The concept of organization ID to maintain the uniqueness among both indexing databases was implemented accordingly. In last step, on the basis of Scopus organization ID and Web of Science organization name, various bibliometric details, like publication count, citation count and  $h$ -index in both indexing databases were retrieved. Fig. 5.13 gives the overview of the filtration process of organization profiles. Filtered organization profiles (100)

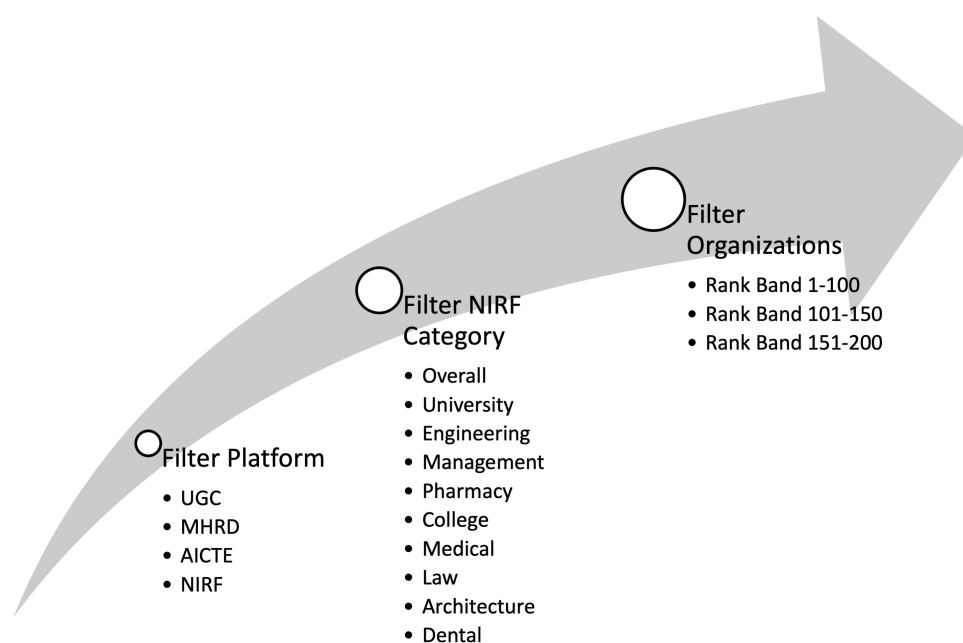


FIGURE 5.13: *Filtration process listing all the steps, from initial selection of platform to final list of organization profiles.*

were categorized on the basis of their types. Fig. 5.14 gives the overview of 100 organizations such as Universities (69), IITs (16), NITs (8), and IEST, IISC & IISER (7).

Universities	IITs	NITs	IIST, IISC & IISER
• 69	• 16	• 8	• 7

FIGURE 5.14: *Details of 100 organizations on the basis of their type.*

### 5.2.1 Number of publications

Fig. 5.15 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of publications of 100 organizations. It is observed that the highest number of publications among different databases are from IITs and the lowest number of publications are from NITs. Conflate reported that the number of publications among different databases is varying between Scopus and the Web of Science across all entities. In all entities, Conflate reported the highest number of publications as compared to Web of Science and the lowest number of articles as compared to Scopus. In Fig. 5.16 during the comparative analysis of a number of publications featured in Scopus, Web of Science, and Conflate, it was identified that the average number of publications in Scopus was 9641, as compared to 8737 in Conflate for 100 organizations. The difference in the average number of publications states that all publications published in Scopus are not considered in Conflate. While comparing the average number of publications on the Web of Science with Conflate, a significant difference may be observed. The average number of publications on the Web of Science is 7971, whereas on Conflate it is 8737.

Table. 5.8 represents the comparative analysis of publications from Scopus, Web of Science, and Conflate for 100 organization profiles among different types.

### 5.2.2 Number of citations

Fig. 5.17 shows the comparison of results generated with Scopus, Web of Science and Conflate on the basis of the number of citations of 100 organizations. The average

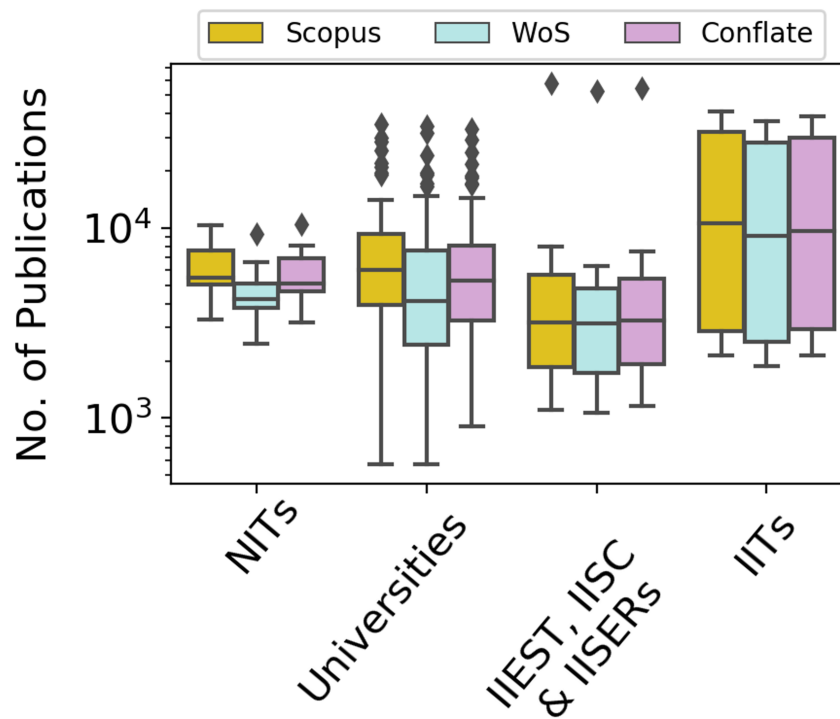


FIGURE 5.15: *A comparison of publications of 100 organizations based on Scopus, Web of Science and Conflate.*

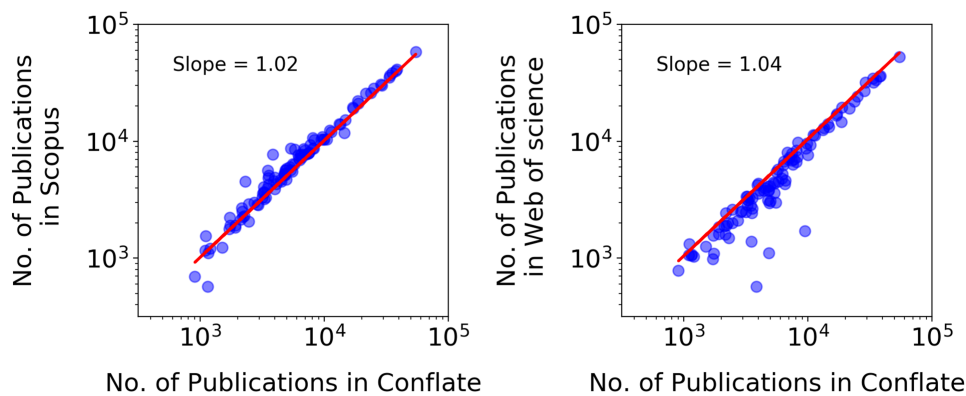


FIGURE 5.16: *Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level.*

number of citations recorded in Scopus is less than in Conflate. This states that although the average number of considered publications is lower in Conflate, the average number of citations is higher. Conflate also reported a significantly higher number of citations as compared to the Web of Science. In Fig. 5.18 during the comparative analysis of the number of citations featured in Scopus, it is observed

Organizations - 100 (Types)	Publications		
	Scopus	Web of Science	Conflate
NITs	50362	39059	47683
Universities	565887	451489	499159
IEST, IISC & IISER	77349	70063	73835
IITs	270495	236547	253042

TABLE 5.8: Comparative analysis of publications - organization level

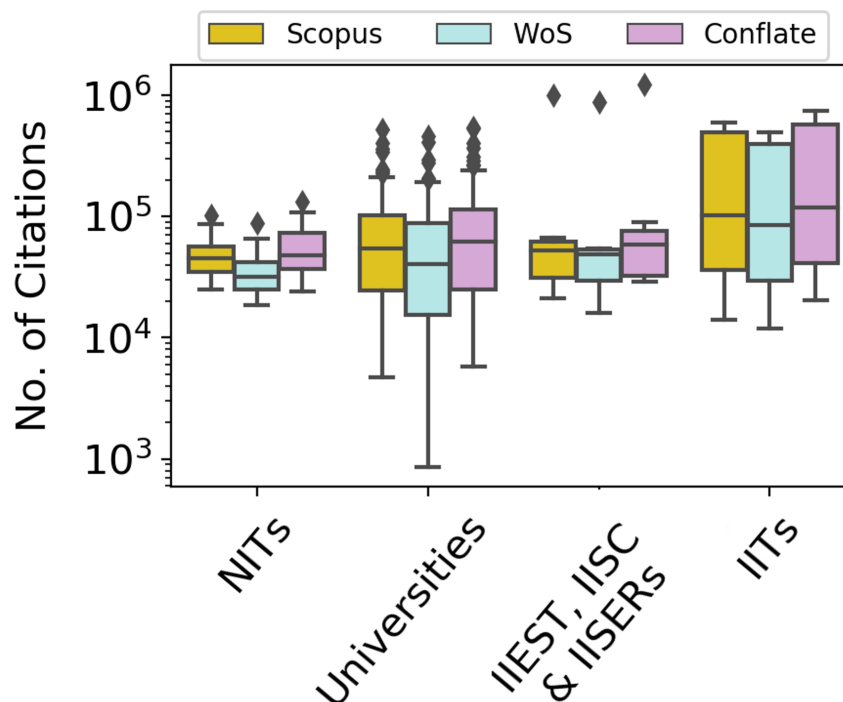


FIGURE 5.17: A comparison of citations of 100 organizations based on Scopus, Web of Science and Conflate.

that the average number of citations received by an organization is 113999, whereas in Conflate it is 134831. The average number of citations published on the Web of Science is 93371 as compared to 134831 in Conflate. The Web of Science has reported the lowest citations, whereas Conflate has reported significantly higher.

Table. 5.9 represents the comparative analysis of citations from Scopus, Web of Science, and Conflate for 100 organization profiles among different types.



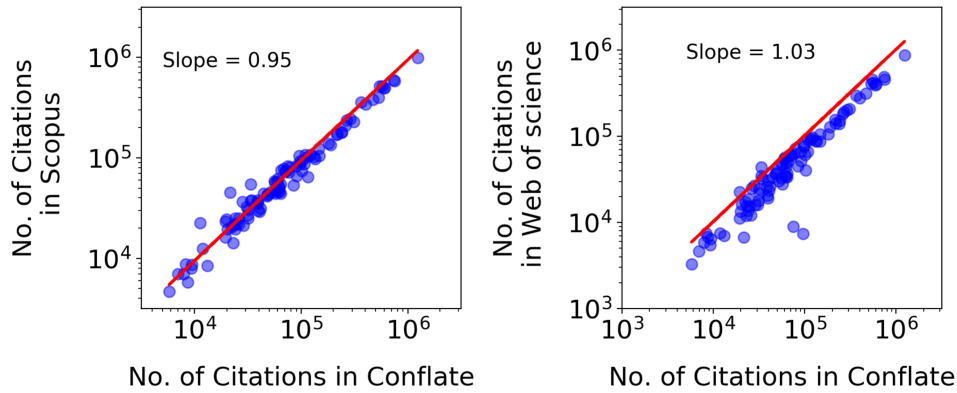


FIGURE 5.18: *Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level.*

Organizations - 100 (Types)	Citations		
	Scopus	Web of Science	Conflate
NITs	416302	318676	492760
Universities	6051897	4917831	6997951
IEST, IISC & IISER	1252987	1107018	1542242
IITs	3678723	2993534	4450159

TABLE 5.9: *Comparative analysis of citations - organization level*

### 5.2.3 Measuring the $h$ -index

Fig. 5.19 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of publications and citations of 100 organizations. For the  $h$ -index of 100 organizations, it was found that Conflate has reported the highest  $h$ -index among both indexing databases. IITs have received the highest  $h$ -index and NITs have received the lowest  $h$ -index among other entities.

Conflate also reported that the results generated are always in between the range of Scopus and Web of Science. Among four entities, it can be observed that IITs have the highest  $h$ -index across multiple databases. It can be stated that different kinds of organizations have different contributions in the field of scientific work. In Fig. 5.20 during the comparative analysis of  $h$ -index featured in Scopus, it is observed that the average  $h$ -index received by an organization is 91, whereas in Conflate it is 100. The average  $h$ -index received in the Web of Science is 82, as compared to 100 in Conflate.

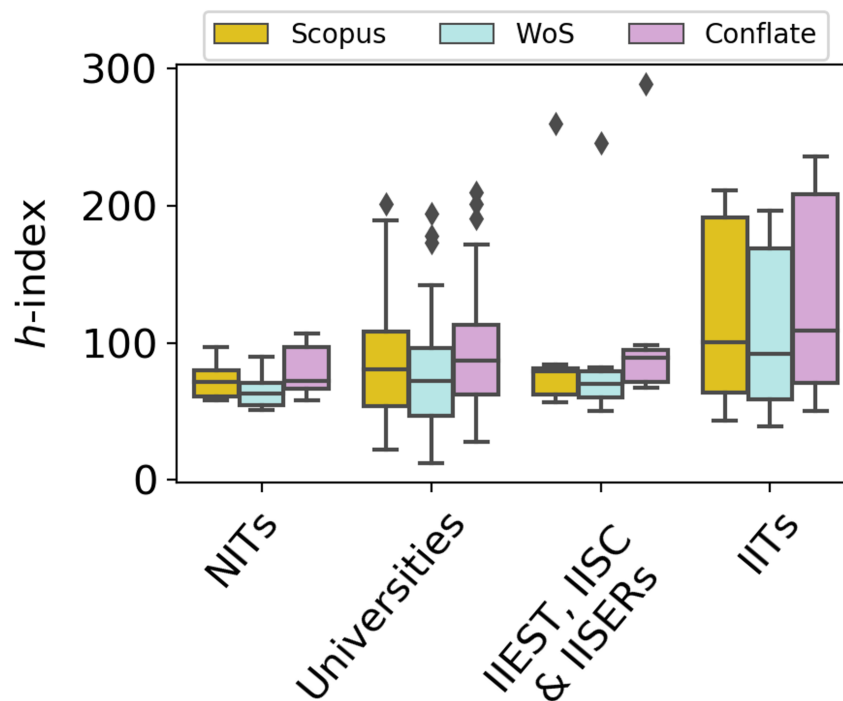


FIGURE 5.19: *A comparison of h-index of 100 organizations based on Scopus, Web of Science and Conflate.*

The Web of Science has reported the lowest average  $h$ -index whereas Conflate has reported the highest. Table. 5.10 represents the comparative analysis of average  $h$ -index from Scopus, Web of Science, and Conflate for 100 organization profiles among different types.

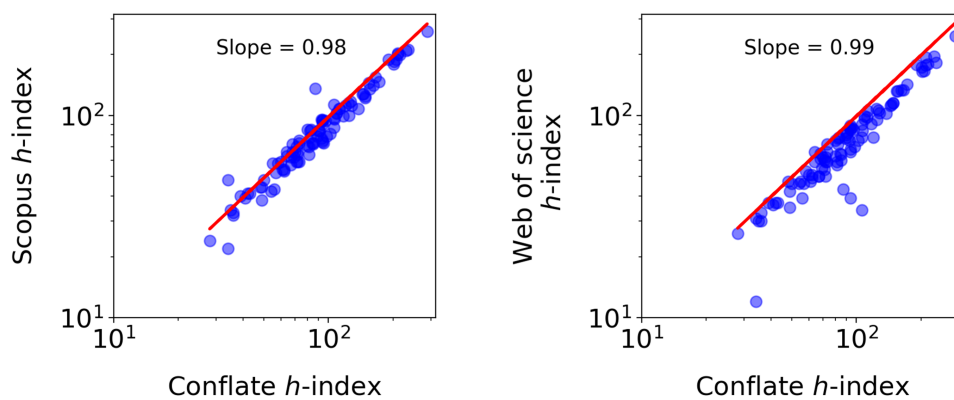


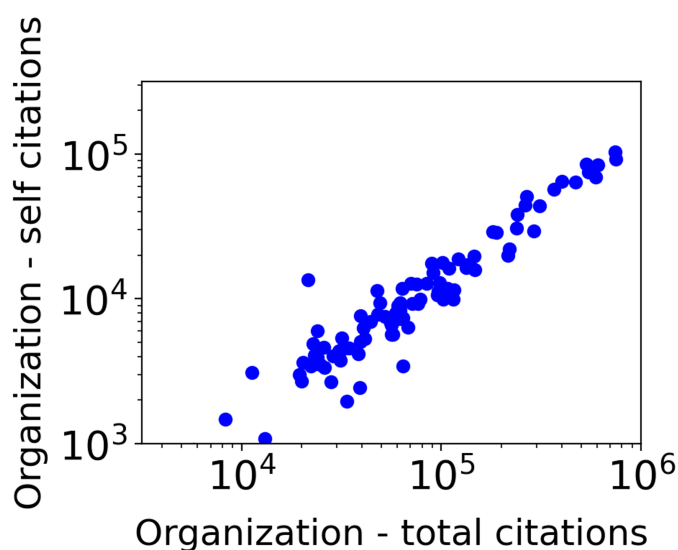
FIGURE 5.20: *Comparative analysis based on h-index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at organization level.*

Organizations - 100 (Types)	Average $h$ -index		
	Scopus	Web of Science	Conflate
NITs	74	66	80
Universities	85	76	93
IEST, IISC & IISER	98	92	111
IITs	122	110	134

TABLE 5.10: *Comparative analysis of average  $h$ -index - organization level*

### 5.2.4 Self-citations vs. total-citations

Comparative analysis of 100 organizations on the basis of their total and self citations in Fig. 5.21 states that average number of total citations for organizations is 134831.12 and average number of self citations for organizations is 18452.23. It can be concluded here that approximately 7.30 citations in the case of organizations are self-citations as compared to total citations. Table. 5.11 represents the comparative

FIGURE 5.21: *A comparison of total and self citations of 100 organizations based on Scopus, Web of Science and Conflate.*

analysis of total citations and self-citations for 100 organizations among different types.

Organizations - 100 (Types)	Self citations	Total citations
NITs	63545	492760
Universities	970979	6997951
IEST, IISC & IISER	201982	1542242
IITs	608717	4450159

TABLE 5.11: *Comparative analysis of self citations vs. total citations - organization level*

### 5.2.5 Repeated-citations vs. total-citations

A comparative analysis of 100 organizations on the basis of their total and repeated citations in Fig. 5.22 states that the average number of total citations for organizations is 134831.12 and the average number of repeated citations for organizations is 45416.84. It can be concluded here that approximately 2.97 citations in the case of organizations are repeated citations as compared to total citations. Table. 5.12

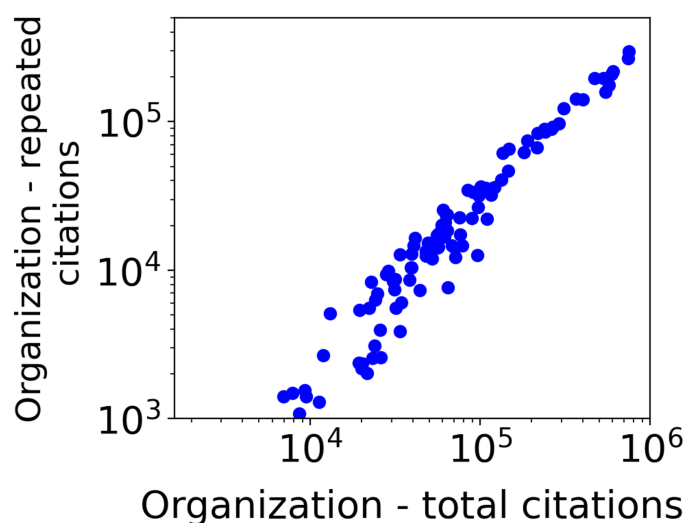


FIGURE 5.22: *A comparison of total and repeated citations of 100 organizations based on Scopus, Web of Science and Conflate.*

represents the comparative analysis of total citations and repeated citations for 100 organizations among different types.

Organizations - 100 (Types)	Repeated citations	Total citations
NITs	136425	492760
Universities	2254465	6997951
IEST, IISC & IISER	559442	1542242
IITs	1591352	4450159

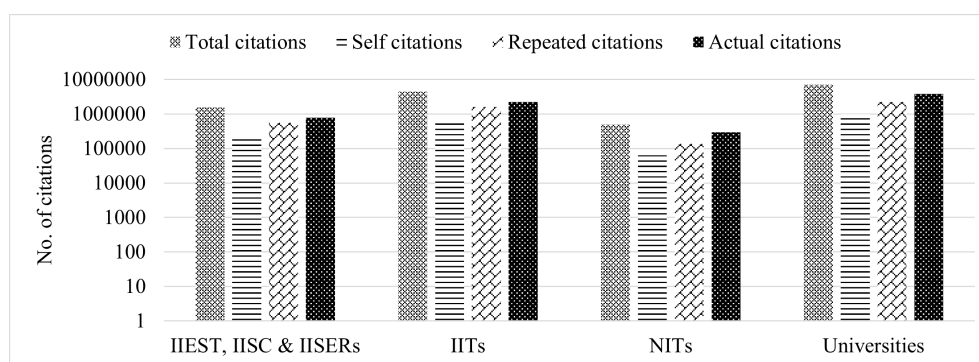
TABLE 5.12: *Comparative analysis of repeated citations vs. total citations - organization level*

Organizations - 100 (Types)	Actual citations	Self citations	Repeated citations	Total citations
NITs	292790	63545	136425	492760
Universities	3772507	970979	2254465	6997951
IEST, IISC & IISER	780818	201982	559442	1542242
IITs	2250090	608717	1591352	4450159

TABLE 5.13: *Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - organization level*

## 5.2.6 Actual-citations vs. total-citations

Comparative analysis of 100 organizations on the basis of their actual, self, repeated and total citations in Fig. 5.23 states that average number of total citations for organizations is 134831.12, average number of repeated citations for organizations is 45416.84, average number of self citations for organizations is 18452.23 and average number of actual citations is 70962.05. Table. 5.13 represents the comparative analysis

FIGURE 5.23: *A comparison of total, self, repeated and actual citations of 100 organizations based on Scopus, Web of Science and Conflate.*

of actual citations, self-citations, repeated citations, and total citations for 400 author profiles among different disciplines.

### 5.2.7 No. of citations vs. average $h$ -index

Fig. 5.24 shows the comparative analysis of citations with  $h$ -index for organizations among multiple indexing databases. During citation level analysis of organizations in Scopus, it is observed that the average number of citations is 113999.09 and the average  $h$ -index earned is 91.09. It can be stated here that average cost of  $h$ -index is 1251.50 per citation.

While analyzing the results of Web of Science, it is observed that average number of citations is 93370.60 against the average  $h$ -index of 81.5 which means that average cost is 1145.65 which is less than the average cost of 1251.50 calculated in Scopus. For Conflate, the average number of citations for authors is 134831.12 against the average  $h$ -index of 99.51, costing around 1354.95 per  $h$ -index as compared to 1145.65 in the Web of Science and 1251.50 in Scopus. Table. 5.14 represents the comparative

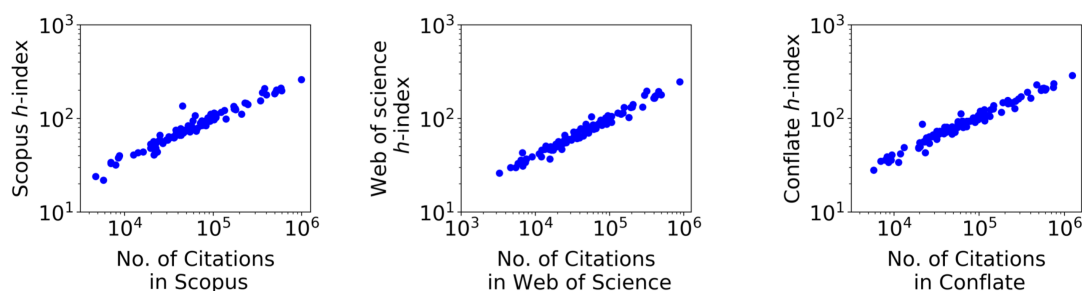


FIGURE 5.24: *A comparison of citations and  $h$ -index of 100 organizations based on Scopus, Web of Science and Conflate.*

analysis of total citations, and average  $h$ -index for 100 organization profiles among different types.

## 5.3 Journal level bibliometrics

The final destination of a quality scientific work is a publication. There are different places where an author can publish his scientific work, like journals, book series, conferences, and proceedings, etc. A scientific work is most often published in a journal. The first is the journal's discipline or subject area of publication, and the

Indexing databases	Scopus		Web of Science		Conflate	
	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index
Organizations - 100 (Types)						
NITs	416302	74	318676	66	492760	80
Universities	6051897	85	4917831	76	6997951	93
IEST, IISC & IISER	1252987	98	1107018	92	1542242	111
IITs	3678723	122	2993534	110	4450159	134

TABLE 5.14: *Comparative analysis of no. of citations vs. average h-index - organization level*

second is its impact factor. There are different disciplines or subject areas like agriculture, accounting, astronomy, computer science, engineering, humanities, medicine, social sciences, etc., in which an author can submit his scientific work.

The impact factor is considered an important bibliometric for the evaluation of a journal. The higher the impact factor, the higher the credibility and quality of scientific work published in that journal. To perform citation analysis at the level of journals, data from both indexing databases, such as Scopus and Web of Science, were retrieved. To maintain the uniqueness of journals in citation analysis, the concept of ISSN was used. All journals indexed in Scopus and Web of Science are maintained by their ISSN numbers. Initially, publications and citations were counted on the basis of ISSN numbers and retrieved [136].

Following that, the discipline and subject details of all journals were added so that citation analysis could be performed on almost all disciplines and subject areas. There were approximately 1195 journals, and a sample of 1000 journals for citation analysis was finalized. Fig. 5.25 gives the overview of the filtration process of journal profiles. Filtered journal profiles (1000) were categorized on the basis of their disciplines. Fig. 5.26 gives the overview of disciplines or subject areas of 1000 journals such as engineering (800), social sciences (119), life sciences (35), sciences (27), and humanities (19).

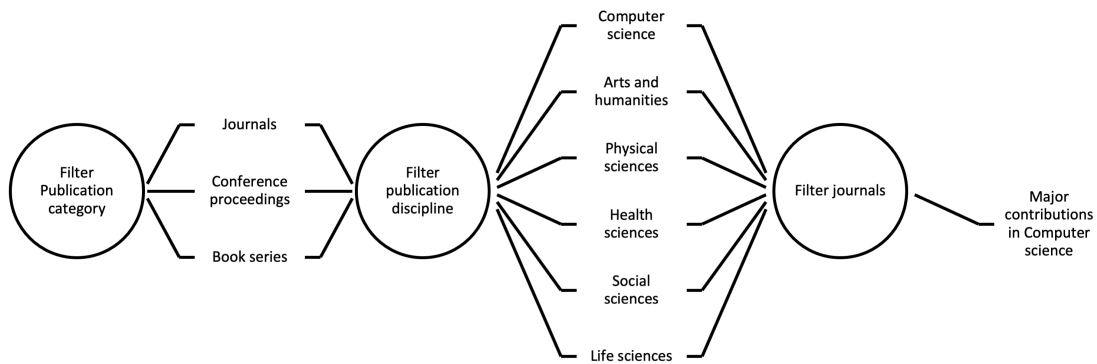


FIGURE 5.25: *Filtration process listing all the steps, from initial selection of journals to final list of journals.*

Engineering	Social Sciences	Life Sciences	Sciences	Humanities
• 800	• 119	• 35	• 27	• 19

FIGURE 5.26: *Details of 1000 journals on the basis of their disciplines or subject areas.*

### 5.3.1 Number of publications

Fig. 5.27 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of publications in 1000 journals. The number of publications observed in the sciences is highest in Scopus and lowest in social sciences. For social sciences, Conflate reported the highest number of publications among Scopus and Web of Science. For the humanities, engineering, and sciences, Conflate has reported a number of publications in the range of Scopus and Web of Science. For life sciences, Conflate has reported almost the same number of publications as compared to Scopus, which is quite less than the Web of Science database. In Fig. 5.28 during the comparative analysis of a number of publications featured in Scopus, Web of Science, and Conflate. While comparing the average number of articles in Scopus (1529) with that of Conflate (1482), it is observed that there is a slight hike in the average number of publications in Scopus. On the other hand, the average number of publications in the Web of Science (1415) as compared to Conflate



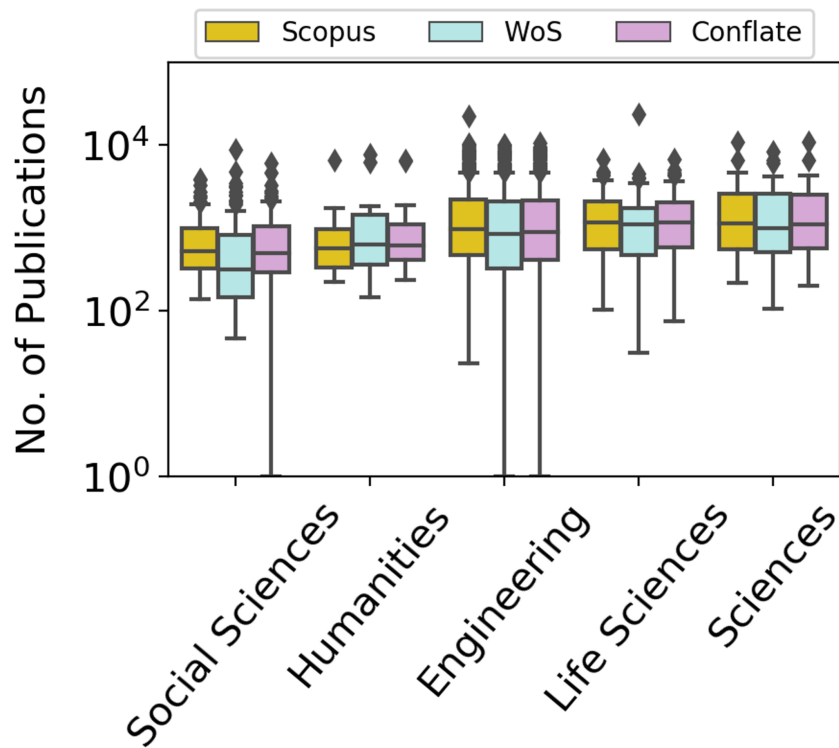


FIGURE 5.27: *A comparison of publications of 1000 journals based on Scopus, Web of Science and Conflate.*

(1482) shows significantly close values.

Table. 5.15 represents the comparative analysis of publications from Scopus, Web of Science, and Conflate for 1000 journal profiles among different disciplines.

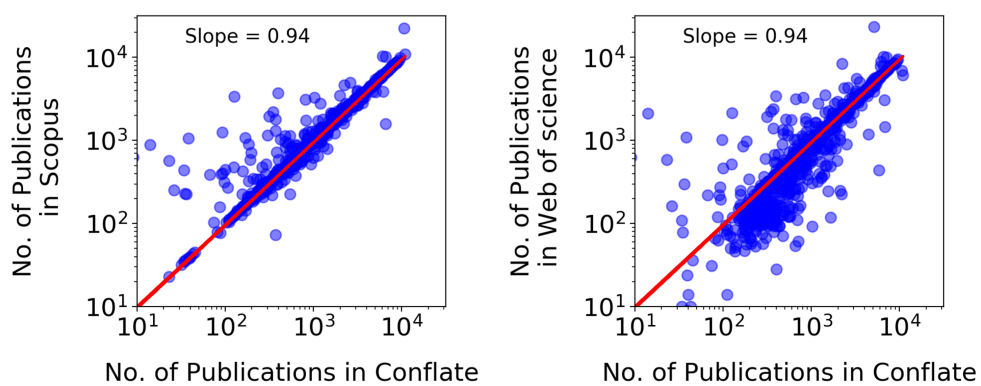


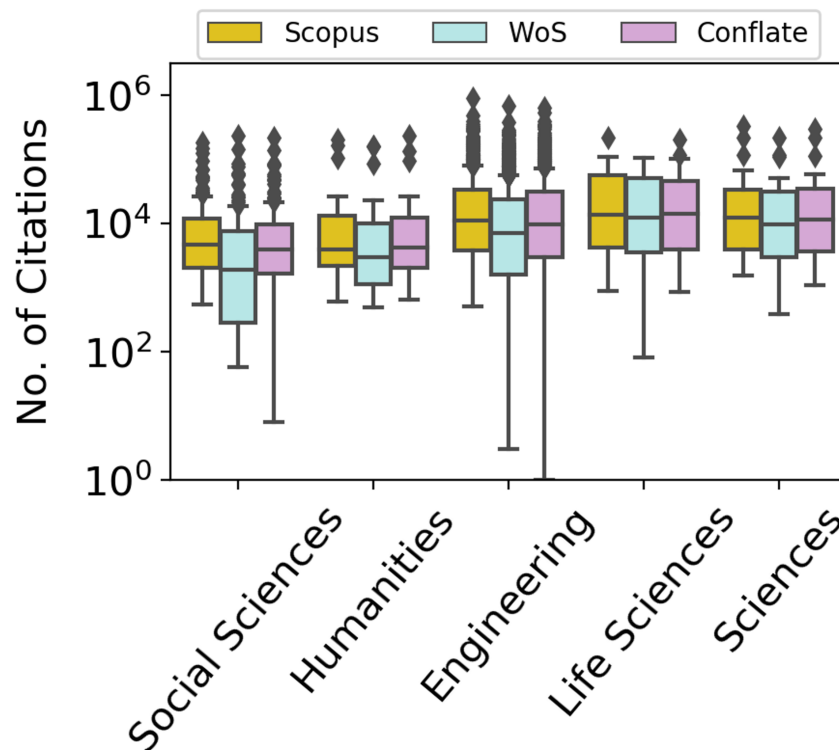
FIGURE 5.28: *Comparative analysis based on number of publications in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level.*

Journals - 1000 (Disciplines)	Publications		
	Scopus	Web of Science	Conflate
Life sciences	60061	68142	60049
Social sciences	94755	86719	98316
Engineering	1298929	1179771	1244492
Sciences	56390	53458	54009
Humanities	18769	27003	24957

TABLE 5.15: *Comparative analysis of publications - journal level*

### 5.3.2 Number of citations

Fig. 5.29 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of citations of 1000 journals. The number of citations reported by Conflate is in between the range of Scopus and the Web of Science for all disciplines, where the sciences are on top and social sciences are at the bottom. In Fig. 5.30 during the comparative analysis of the number of citations

FIGURE 5.29: *A comparison of citations of 1000 journals based on Scopus, Web of Science and Conflate.*

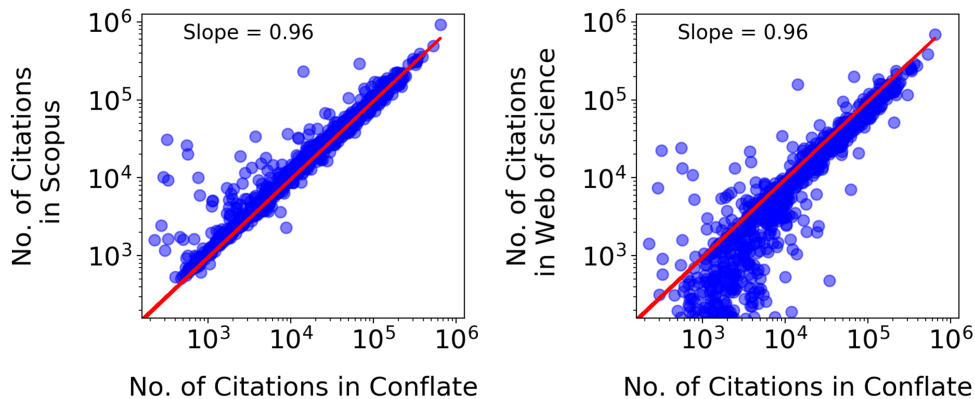
featured in Scopus, it is observed that the average number of citations received by

Journals - 1000 (Disciplines)	Citations		
	Scopus	Web of Science	Conflate
Life sciences	1144780	858220	1130871
Social sciences	1712570	1268836	1569405
Engineering	26468041	19176940	24982597
Sciences	1061160	783108	1030698
Humanities	577741	483357	562547

TABLE 5.16: *Comparative analysis of citations - journal level*

a journal is 30964, whereas in Conflate it is 29276. The average number of citations published on the Web of Science is 22570 as compared to 29276 in Conflate. Conflate clearly states that there is more scope for consideration of citations as compared to citations considered by the Web of Science.

Table. 5.16 represents the comparative analysis of citations from Scopus, Web of Science, and Conflate for 1000 journal profiles among different disciplines.

FIGURE 5.30: *Comparative analysis based on number of citations in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level.*

### 5.3.3 Measuring the $h$ -index

Fig. 5.31 shows the comparison of results generated with Scopus, Web of Science, and Conflate on the basis of the number of publications and citations of 1000 journals. Conflate's  $h$ -index for 1000 journals is the same as Scopus's for humanities, sciences, and life sciences. For social sciences and engineering, it is in between the range of Scopus and the Web of Science. The lowest  $h$ -index is reported by the Web of Science

for social sciences and the highest by Scopus for life sciences. In Fig. 5.32 during the

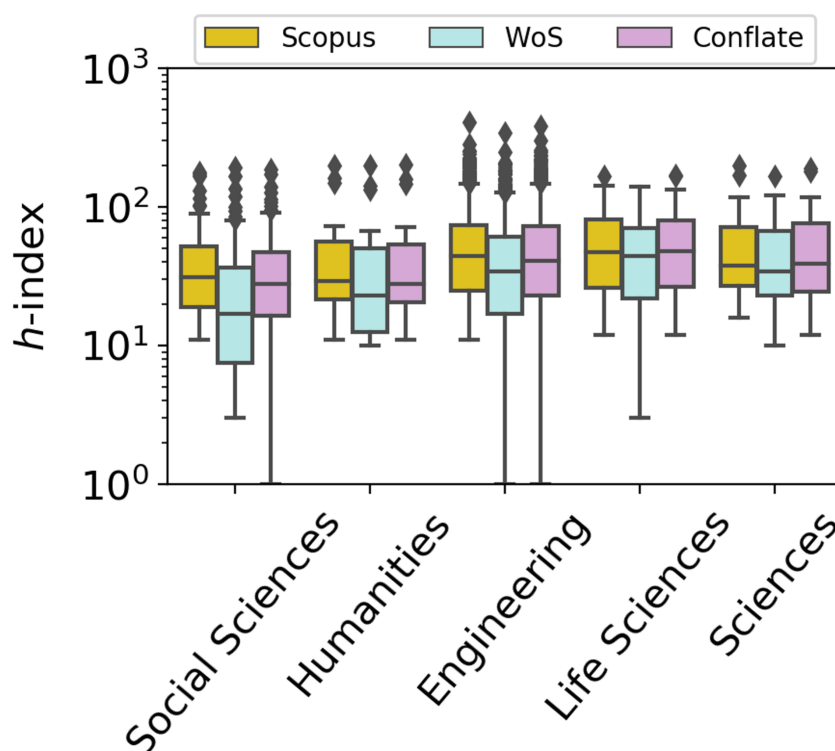


FIGURE 5.31: *A comparison of  $h$ -index of 1000 journals based on Scopus, Web of Science and Conflate.*

comparative analysis of  $h$ -index featured in Scopus, it is observed that the average  $h$ -index received by an journal is 56, whereas in Conflate it is 53. The average  $h$ -index received in the Web of Science is 44 as compared to 53 in Conflate. Web of Science has reported lowest average  $h$ -index whereas Conflate has reported in the range of Scopus and Web of Science. Table. 5.17 represents the comparative analysis of the average  $h$ -index from Scopus, Web of Science, and Conflate for 1000 journal profiles among different disciplines.

#### 5.3.4 Self-citations vs. total-citations

A comparative analysis of 1000 journals on the basis of their total and self citations in Fig. 5.33 states that the average number of total citations for journals is 29276.12 and the average number of self citations for journals is 2254.19. It can be concluded here that approximately 12.99 citations in the case of journals are self-citations,

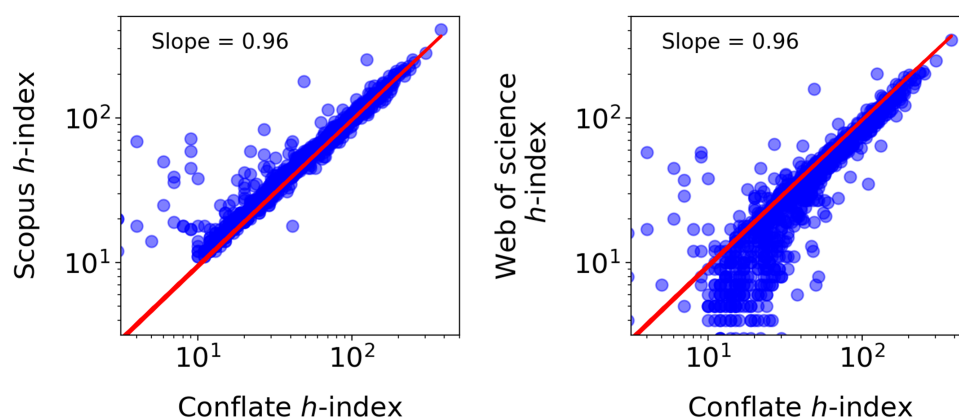


FIGURE 5.32: *Comparative analysis based on  $h$ -index in Scopus (left panel) and Web of Science (right panel) with unified informetrics at journal level.*

Journals - 1000 (Disciplines)	Average $h$ -index		
	Scopus	Web of Science	Conflate
Life sciences	59	50	59
Social sciences	42	30	38
Engineering	58	46	55
Sciences	56	49	56
Humanities	54	46	53

TABLE 5.17: *Comparative analysis of average  $h$ -index - journal level*

and it is quite a high number of self-citations observed for journals. Table. 5.18 rep-

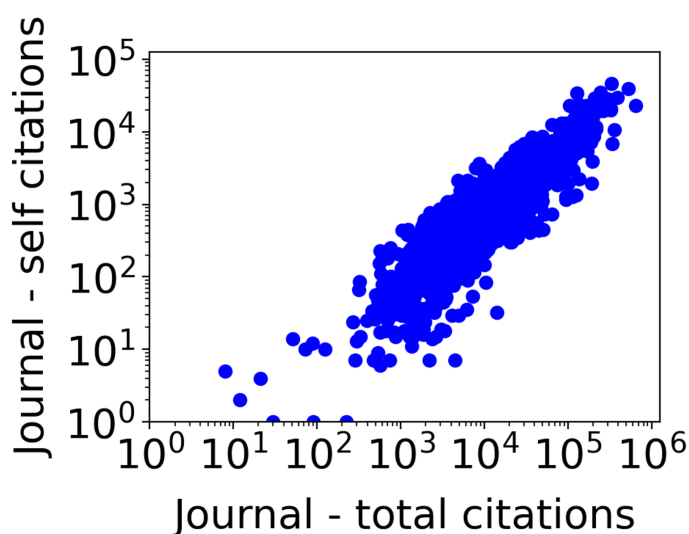


FIGURE 5.33: *A comparison of total and self citations of 1000 journals based on Scopus, Web of Science and Conflate.*

Journals - 1000 (Disciplines)	Self citations	Total citations
Life sciences	65894	1130871
Social sciences	124538	1569405
Engineering	1915771	24982597
Sciences	105973	1030698
Humanities	42012	562547

TABLE 5.18: *Comparative analysis of self citations vs. total citations - journal level*

resents the comparative analysis of total citations and self citations for 1000 journal profiles among different disciplines.

### 5.3.5 Repeated-citations vs. total-citations

Comparative analysis of 1000 journals on the basis of their total and repeated citations in Fig. 5.34 states that the average number of total citations for journals is 29276.12 and the average number of repeated citations for journals is 10429.15. It can be concluded here that approximately 2.981 citations in the case of journals are repeated citations. Table. 5.19 represents the comparative analysis of total citations

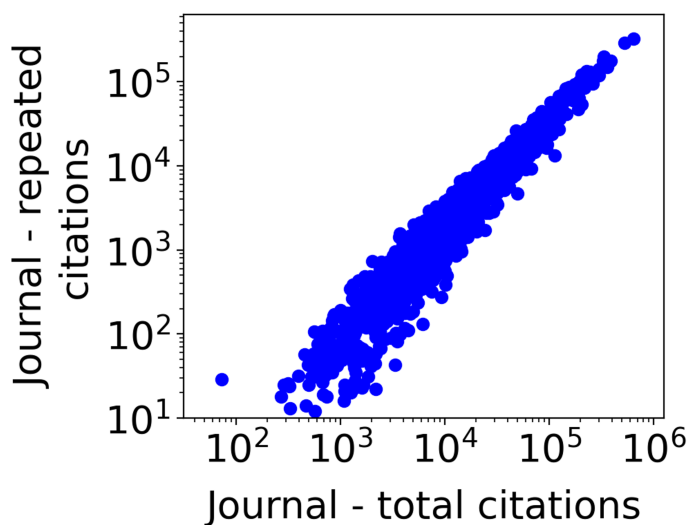


FIGURE 5.34: *A comparison of total and repeated citations of 1000 journals based on Scopus, Web of Science and Conflate.*

and repeated citations for 1000 journal profiles among different disciplines.

<b>Journals - 1000 (Disciplines)</b>	<b>Repeated citations</b>	<b>Total citations</b>
<b>Life sciences</b>	335746	1130871
<b>Social sciences</b>	522019	1569405
<b>Engineering</b>	8951933	24982597
<b>Sciences</b>	401717	1030698
<b>Humanities</b>	217735	562547

TABLE 5.19: *Comparative analysis of repeated citations vs. total citations - journal level*

### 5.3.6 Actual-citations vs. total-citations

Comparative analysis of 1000 journals on the basis of their actual, self, repeated and total citations in Fig. 5.35 states that the average number of total citations for journals is 29276.12, the average number of repeated citations for journals is 10429.15, the average number of self-citations for journals is 2254.19, and the average number of actual citations is 16592.78. Table. 5.20 represents the comparative analysis of actual

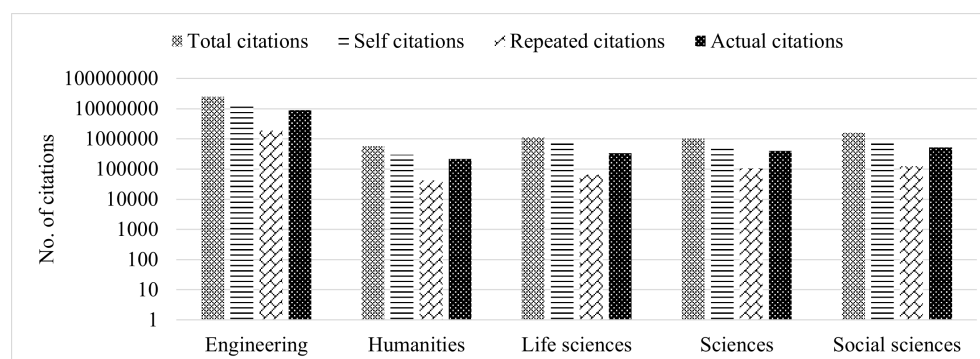


FIGURE 5.35: *A comparison of total, self, repeated and actual citations of 1000 journals based on Scopus, Web of Science and Conflate.*

citations, self citations, repeated citations and total citations for 1000 journal profiles among different disciplines.

### 5.3.7 No. of citations vs. average $h$ -index

Fig. 5.36 shows the comparative analysis of citations with  $h$ -index for journals among multiple indexing databases. During citation level analysis of journals in Scopus, it is observed that the average number of citations is 30964.30 and the average

<b>Journals - 1000 (Disciplines)</b>	<b>Actual citations</b>	<b>Self citations</b>	<b>Repeated citations</b>	<b>Total citations</b>
<b>Life sciences</b>	729231	65894	335746	1130871
<b>Social sciences</b>	922848	124538	522019	1569405
<b>Engineering</b>	14114893	1915771	8951933	24982597
<b>Sciences</b>	523008	105973	401717	1030698
<b>Humanities</b>	302800	42012	217735	562547

TABLE 5.20: *Comparative analysis of actual-citations vs. self-citations vs. repeated citations vs. total-citations - journal level*

$h$ -index earned is 55.71. It can be stated here that the average cost of  $h$ -index is 555.81 per citation.

While analyzing the results of Web of Science, it is observed that average number of citations is 22570.46 against the average  $h$ -index of 43.87 which means that average cost is 514.49 which is less than the average cost of 555.81 calculated in Scopus. For Conflate, the average number of citations for journals is 29276.12 against the average  $h$ -index of 53.30, costing around 549.27 per  $h$ -index as compared to 514.49 in the Web of Science and 555.81 in Scopus. Table. 5.21 represents the comparative

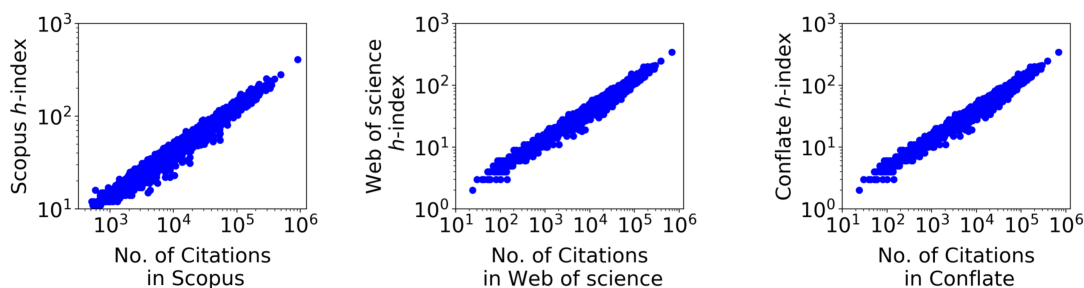


FIGURE 5.36: *A comparison of citations and  $h$ -index of 1000 journals based on Scopus, Web of Science and Conflate.*

analysis of total citations, and average  $h$ -index for 1000 journal profiles among different disciplines.

### 5.3.8 Measuring the impact factor

The quality of the journal is determined by its impact factor. Authors publishing their scientific work in different journals always look towards the impact factor. It is



Indexing databases	Scopus		Web of Science		Conflate	
	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index	Total citations	Avg. $h$ index
<b>Journals - 1000 (Disciplines)</b>						
<b>Life sciences</b>	1144780	59	858220	50	1130871	60
<b>Social sciences</b>	1712570	42	1268836	30	1569405	38
<b>Engineering</b>	26468041	58	19176940	46	24982597	55
<b>Sciences</b>	1061160	56	783108	49	1030698	56
<b>Humanities</b>	577741	54	483357	46	562547	53

TABLE 5.21: *Comparative analysis of no. of citations vs. average h-index - journal level*

considered an important and valuable bibliometric indicator for the evaluation and potential of a journal [137]. Despite other available bibliometric indicators, the impact factor is mostly used due to its features and potential [138].

Results generated by citation analysis performed on Scopus and Web of Science were used. Analyzed results were presented in the form of Conflate and Conflate data set is utilized for the calculation of impact factor for 2018. Initial filtration was done on the basis of publication category, then on the basis of publication discipline, and lastly, on the basis of journal outcome in the form of impact factor. To start with, data from 1000 journals of different disciplines was utilized and a filtration was applied as shown in Fig. 5.37 for the final count of 746 journals. Filtered journal

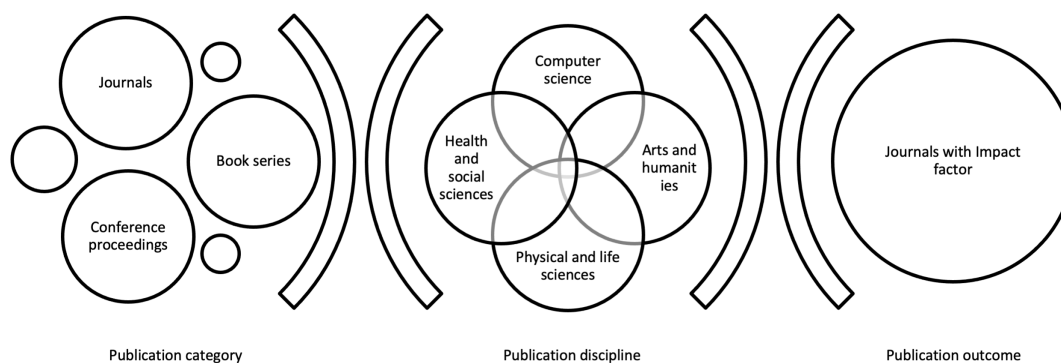


FIGURE 5.37: *Filtration process listing all the steps, from initial selection of journals to final list of journals for impact factor calculation.*

profiles (746) were categorized on the basis of their disciplines. Fig. 5.38 gives the

overview of disciplines or subject areas of 746 journals such as engineering (617), social sciences (64), life sciences (28), sciences (25), and humanities (12). During the citation

Engineering	Social Sciences	Life Sciences	Sciences	Humanities
• 617	• 64	• 28	• 25	• 12

FIGURE 5.38: *Details of 746 journals on the basis of their disciplines or subject areas.*

analysis, it was observed that the average impact factor for 746 journals was 2.41 as per citation reports and in Conflate it was 3.83. The minimum impact factor observed in citation reports was 0.204, and the maximum impact factor was 22.97. In Conflate, the minimum impact factor was 0.38 and the maximum was 15.36, which shows a significant increase in terms of minimum impact factor but a significant decrease in terms of maximum impact factor (see Fig. 5.39). Fig. 5.40 shows that the merging of

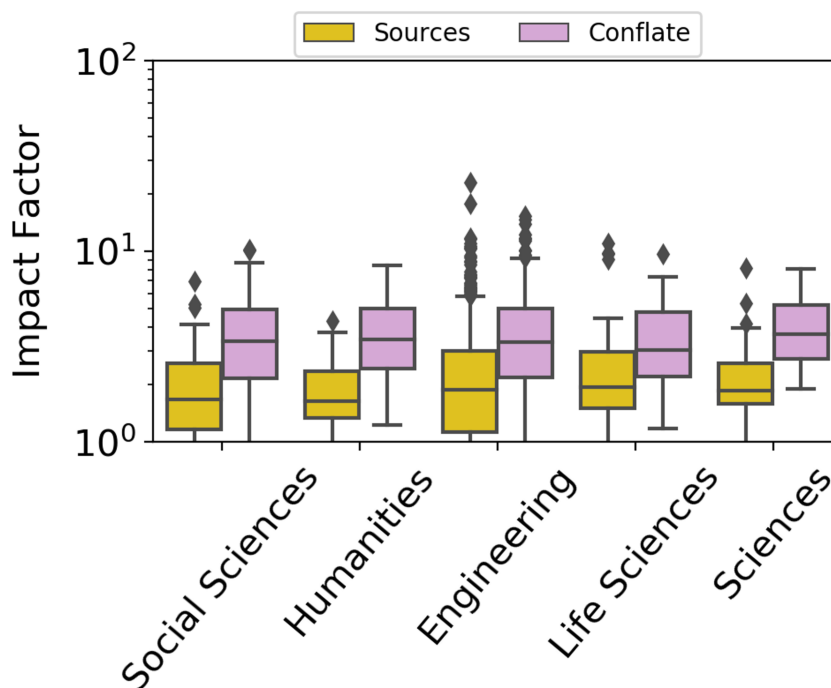


FIGURE 5.39: *A comparison of impact factor of 746 journals based on sources (Scopus and Web of Science together) and Conflate.*

Scopus and the Web of Science for citation analysis has provided a new dimension to

evaluate the research outcomes of different identities. Conflate based impact factor generated as a result set has a minimum value of 0.38 and a maximum value of 15.36, whereas the traditional approach has a minimum value of 0.20 and a maximum value of 22.97 as an impact factor of journals.

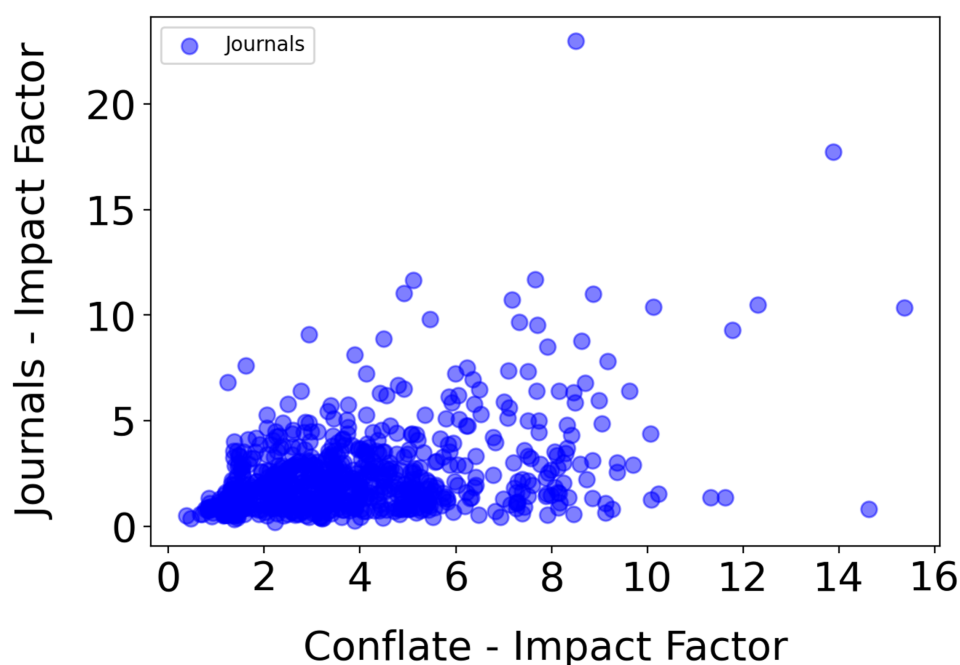


FIGURE 5.40: *Comparison of impact factor on the basis of results generated by Conflate and sources from Scopus and Web of Science.*

## 5.4 Discussion and summary

At the beginning of this chapter, unified informetrics to be performed on three entities categorized as authors, organizations and journals is discussed. Then, the sources of three entities are mentioned, with the idea behind the consideration of these three entities to perform citation analysis. Later on, details on three data sets such as Scopus, Web of Science, and Conflate are discussed.

To calculate unified informetrics for three entities, complete analysis was divided into seven parameters. The first parameter was associated with publications, second parameter was associated with citations; the third parameter was associated with

measuring of  $h$ -index; the fourth, fifth, sixth, and seventh parameters were associated with citations as self citations, repeated citations, total citations, and actual citations. In the third category of journals, citation analysis based on the impact factor of journals is presented.

Significant variation was observed in the results of Conflate. Results were analyzed using both indexing databases collectively as well as individually. In some of the scenarios, it was observed that Conflate has exceeded the values when compared with both indexing databases. In other scenarios, Conflate has presented the values within the range of Scopus and Web of Science. In very few scenarios, it was observed that Conflate had produced the same results as either Scopus or Web of Science.

Finally, after completing citation analysis with two indexing databases, it can be stated that this approach has given a new meaning to citation analysis. In today's fast growing world, when there are multiple indexing databases in the market and new ones are also approaching, this approach or measure can be used as an alternative method to index the scientific work published in different indexing databases and further to give recognition to their authors, organizations and journals.

In the next chapter, the concept of distributed ledger technology in the context of the publication industry is discussed in the form of mapping, consensus, and its implementation.

---

---

## CHAPTER 6

---

# Distributed ledger technology based implementation of Conflate

In this chapter, the concept of distributed ledger technology (DLT) in the context of its analogy, mapping, implementation, and consensus mechanisms is discussed. The question of interest is, how the unified informetrics [139] is mapped with distributed ledger technology and why the interest in the adoption of distributed ledger technology in various industries has increased. So, a mechanism has been derived to bind the research publication industry and distributed ledger technology together. Further, uniqueness among entities such as authors, organizations and journals is maintained. In the end, an algorithm is proposed to implement a consensus mechanism named “Proof of bibliometric indicators (PBI)” to generate a decentralized application based on distributed ledger technology where authors, organizations and journals can keep track of informetrics.

## 6.1 Analogy and consensus for applying DLT

Since its inception, blockchain deployment has been mostly experimental [140]. In 2008, Satoshi Nakamoto (as a man or a group) proposed an idea using blockchain to make internet payments directly from one party to another without the involvement of financial institutions [141]. Since then, blockchain technology has gotten a lot of attention in the financial industry, paving the way for it to extend to other industries. This approach has resulted in a lot of literature on the blockchain. Various authors have emphasized the possible non-financial applications that have sparked interest in blockchain around the world [142, 143]. With capabilities including transaction validation, entry protection, record preservation, immutability, decentralization, consensus, and speedier settlement, blockchain has been viewed as a disruptive combustion engine with the potential to transform business processes and digitize the transaction workflow. The response of blockchain adoption has defined the concept's first manifestation as a Bitcoin, which was inextricably linked to blockchain technology [144–146].

Since then, the authors have shown a keen interest in blockchain literature, particularly in terms of features, concepts, applications, adoption issues, and chances to impact the technology's potential [147]. There has also been a lot of literature predicting new industry-driven solutions [148] and promising huge business gains from blockchain technology. Innovations, new trends, and corporate confidence have all been noted in the literature about blockchain adoption. Various scholars have also looked at blockchain as a collaboration that is generating more momentum and solving specific problems for growing economies in various domains [149, 150].

The introduction of blockchain technology in various fields such as banking, health care, agriculture, manufacturing, and others has caused technological waves [151]. The blockchain literature is growing day by day, thanks to the expanded scope of blockchain and high-potential developing technologies such as the Internet of Things (IoT), smart contracts, security, smart properties, and supply chains, etc. [152, 153]. The blockchain's reach has been widely expanded to a range of applications, according to several bibliometric evaluations. Blockchain implementers have highlighted

the rise of blockchain as a breakthrough for several potential industrial models [154]. Recorded and validated transactions, synchronized involvements, time-stamped entries, and transparent systems have contributed to blockchain becoming a more efficient and controlled product around the world. Systematic reviews of the literature have found appropriate and promising approaches to integrate blockchain with real-world data. Such integration and transformation of literature has demonstrated that blockchain has enormous potential, with enormous opportunities still to be discovered around the world [155].

### **6.1.1 Mapping of distributed ledger technology with research publications**

Every contributor to a large system believes and accesses the distributed ledger data bank independently. The allotment is exclusive, and each node holds and constructs records independently [156]. A block serves as a register to record the transaction information in encrypted form and is identified by the hash based on the information stored in the previous block and current block. It is thus a stable collection of records that cannot be modified or erased once engraved. In a direct-sequential order, the blocks are added to the chain. The chain structure permanently time stamps and stores value exchanges, making it impossible for anybody to alter the ledger [157]. Each block's record contains at least one transaction; however, a single block can include numerous effective transactions. Every contributing node's transaction record (ledger entry) is linked to previous transactions and is consistent. Every ledger record can be traced back to the beginning of time and remodeled [158].

The details of how the terminologies in DLT are mapped with academic research industry are shown in Fig. 6.1. This mapping methodology serves as the foundation for terminology used in the research publishing sector and in DLT. This would aid the system in determining the best combination of parts to use when constructing designs.

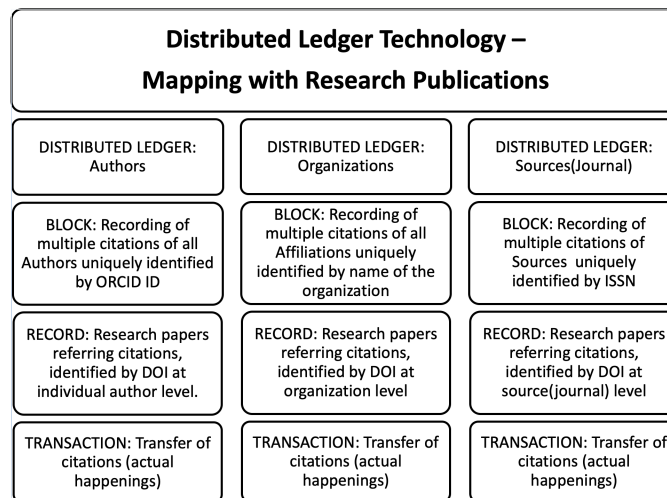


FIGURE 6.1: Schematic representation of DLT mapping with research publications.

## 6.1.2 Proof of bibliometric indicators (PBI) - consensus mechanism

Here we have presented a detailed description of the DLT-based system designed for unified informetrics and consensus mechanism.

### 6.1.2.1 Design of DLT based system for unified informetrics

The system model of proof of bibliometric indicators (PBI) proposed in Fig. 6.2 is divided into two stages:

- *Generation of unified informetrics (Stage-1)*: To access the list of publications, citations, and  $h$ -index from indexing databases, author credentials such as OrCID ID will be necessary. Unified informetrics will be developed using several indexing databases.
- *Implementation of proof of bibliometric indicators (Stage-2)*: The author will be shown the generated informetrics in order to obtain his permission to publish unified informetrics in a block. If the author selects "Yes," the system will seek the author's permission to display more information. If the author accepts, he or she will be given the option of choosing different informetrics; otherwise, unified informetrics will be transmitted for block generation.



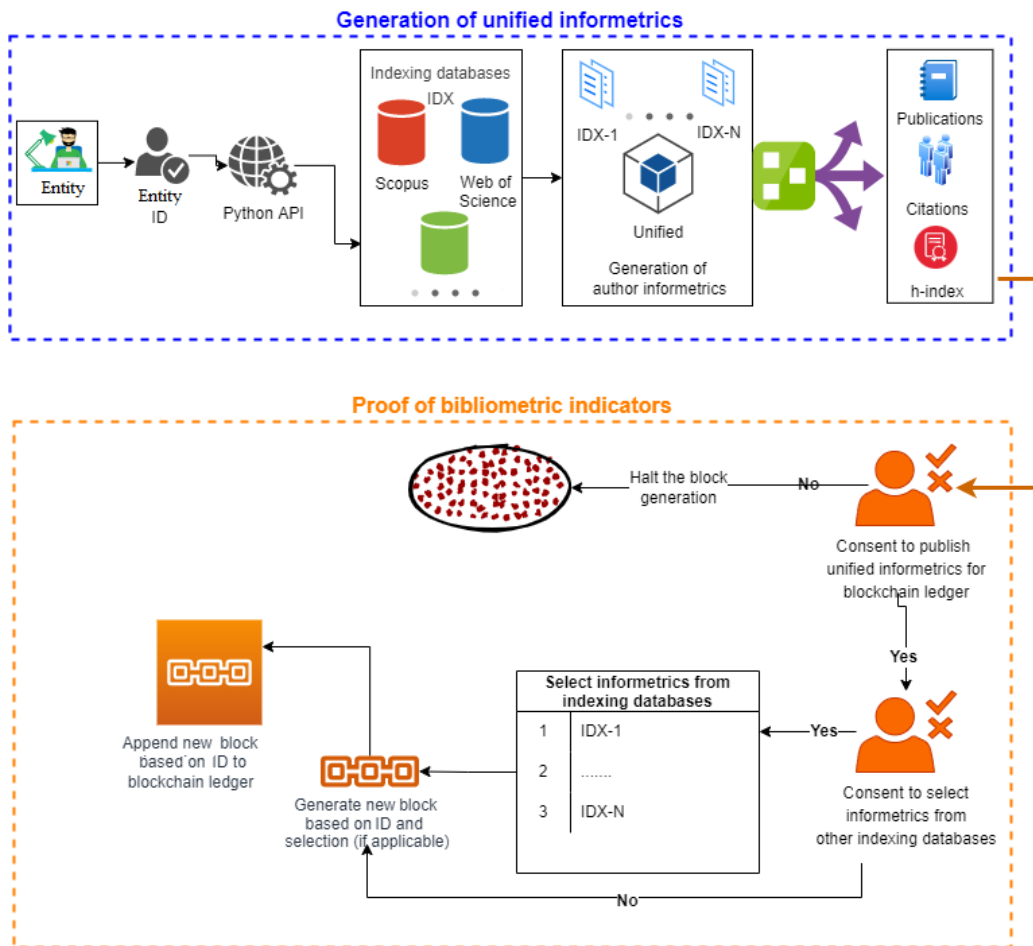


FIGURE 6.2: Representation of system model of PBI.

### 6.1.2.2 Design of PBI consensus mechanism

Algorithm 4 presents the step-by-step procedure for PBI consensus. Entity identifiers such as an author’s ORCID, an organization’s name, and a journal’s ISSN are used to extract bibliometrics. System will access entity’s required bibliometrics from available indexing databases with the list of publications, citations and  $h$ -index information. By using extracted data, filtration, and weight assignment, a unified informetrics will be generated.

In the next step, the entity’s consent to publish unified informetrics to a distributed ledger will be taken. If the entity replies “no”, the process is halted. If the entity replies “yes,” the entity will be asked to select informetrics from multiple indexing databases as well. If the entity replies “yes”, a new block will be generated

and added to the entity ledger.

---

**Algorithm 4** Proof of bibliometric indicators (PBI)

---

**Require:** Entity identifier

**Ensure:** Block generation of informetrics in blockchain

```

1: for each entity do
2:    $[AI_i] \in IDX_i$ , where  $i = 1, \dots, N, N > 0$ 
3:    $\triangleright$  /*  $[AI_i]$  is list of entity's informetrics like #Publications, #Citations,
   #h-index, etc. in indexing databases  $IDX_i$ . */
4:    $\triangleright$  /* Generate unified informetrics  $UI$ , where  $UI$  is generated from unique doi's
   of  $N$  indexing databases. */
5:   while  $N > 0$  do
6:      $UI = IDX_1 \cup IDX_2 \cup \dots \cup IDX_N$ 
7:   end while
8:    $\triangleright$  /* Entity consent to generate block */
9:   if Entity Consent to display  $UI$  is "YES": then
10:    if Entity consent to display other  $AI$  is "YES" then
11:      Select  $AI$  and generate block
12:    else
13:      Generate block based on  $UI$ 
14:    end if
15:  else
16:    Halt the block generation
17:  end if
18: end for

```

---

## 6.2 Implementation details

Here we have presented the detailed description of the DLT-based system implemented for unified informetrics with input/output description, entity registration, and block generation.

### 6.2.1 Input and output demonstration of informetrics in blockchain

The admin registers identity information such as author, organization, and journal, using a decentralized application built with Truffle and Ganache. Truffle provides a development environment for smart contracts as well as a framework for testing them on the Ethereum Virtual Machine (EVM). Ganache is a local blockchain that is used

to run tests and deploy smart contracts. The entity name or ID, number of publications, total citations, self-citations, repeated citations,  $h$ -index, and actual citations are maintained for each entity. Informetrics of entities, uniquely recognized by entity name or ID, are kept in blocks of the blockchain in the implemented informetrics system.

## 6.2.2 Entity registration using Identifiers

Here we have presented the detailed description of the entity registration at the author, organization, and journal level for DLT implementation.

### 6.2.2.1 Author level bibliometrics

The admin interface is shown in Fig. 6.3 (a), where a new author is registered. The author will see the author's address, Orcid ID, and blockchain smart contract address. The sample of author citation transactions of a smart contract deployed on Ganache is shown in Fig. 6.3 (b). The admin's details from Fig. 6.3 (a) are displayed, and the admin is prompted to proceed with author credentials and transaction data on the server. The system generated author address and smart contract address information are shown in Fig. 6.3 (c), indicating that the author has been successfully registered on the smart contract based decentralized application.

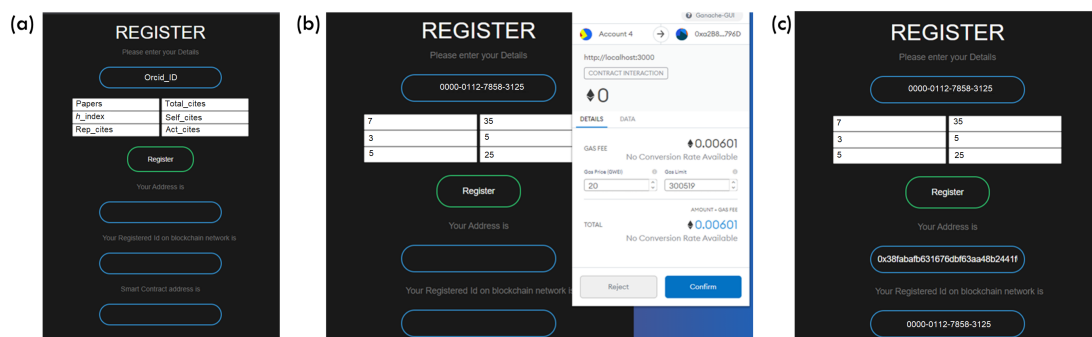


FIGURE 6.3: *Author registration process for (a) user interface (b) transaction (c) registration confirmation.*

### 6.2.2.2 Organization level bibliometrics

The admin interface is shown in Fig. 6.4 (a), where a new organization is registered. The organization will see its address, name, and blockchain smart contract address. The sample of organization citation transactions of a smart contract deployed on Ganache is shown in Fig. 6.4 (b). The admin's details from Fig. 6.4 (a) are displayed, and the admin is prompted to proceed with organization credentials and transaction data on the server. The system generated organization address and smart contract address information are shown in Fig. 6.4 (c), indicating that the organization has been successfully registered on the smart contract based decentralized application.

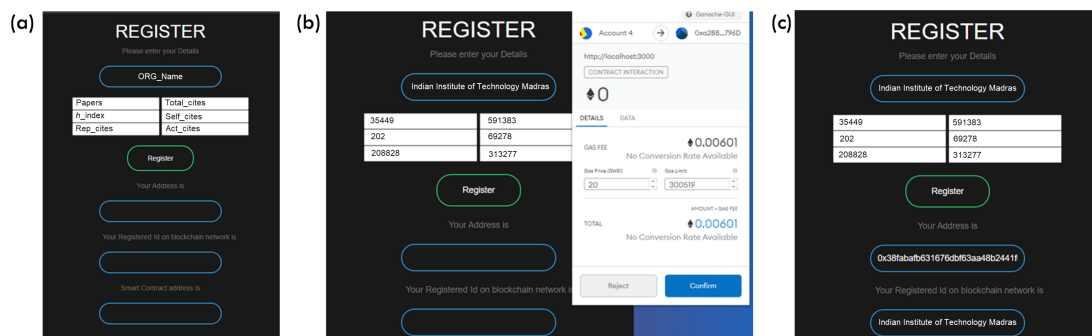


FIGURE 6.4: *Organization registration process for (a) user interface (b) transaction (c) registration confirmation.*

### 6.2.2.3 Journal level bibliometrics

The admin interface is shown in Fig. 6.5 (a), where a new journal is registered. The journal will see its address, ISSN, and blockchain smart contract address. The sample of journal citation transactions of a smart contract deployed on Ganache is shown in Fig. 6.5 (b). The admin's details from Fig. 6.5 (a) are displayed, and the admin is prompted to proceed with journal credentials and transaction data on the server. The system generated journal address and smart contract address information are shown in Fig. 6.5 (c), indicating that the journal has been successfully registered on the smart contract based decentralized application.

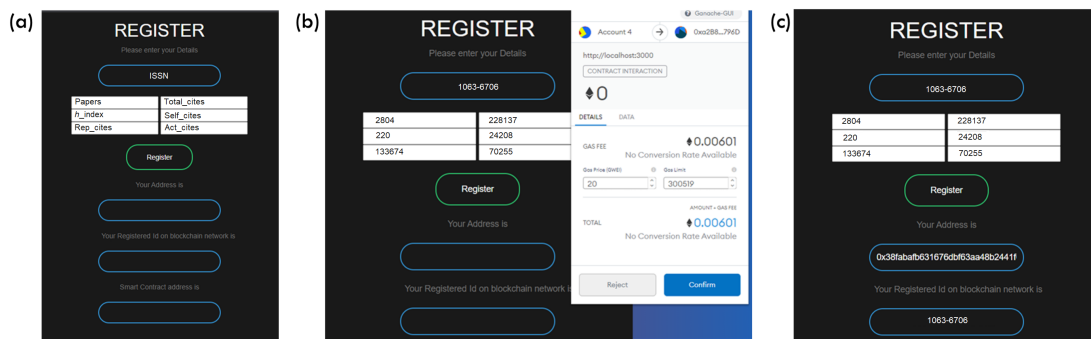


FIGURE 6.5: *Journal registration process for (a) user interface (b) transaction (c) registration confirmation.*

### 6.2.3 Block and transaction generation and confirmation process for all entities

A block indicating the transaction is formed in the blockchain when the entity's information is entered in the decentralized application developed for the implemented informetric system, as seen below (Fig. 6.6): When a transaction is registered in the

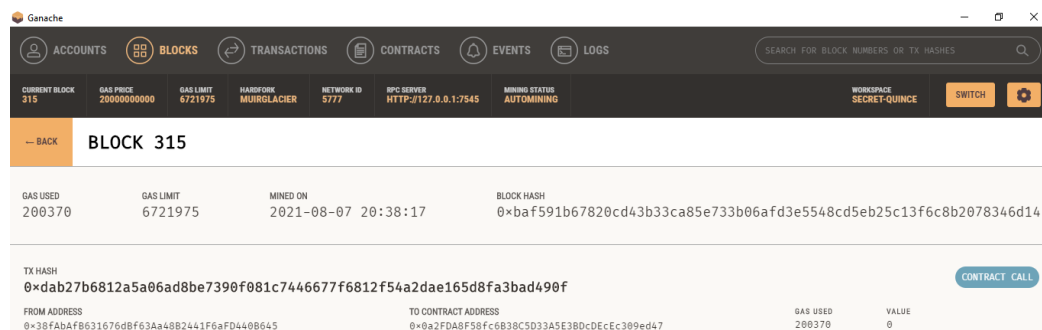
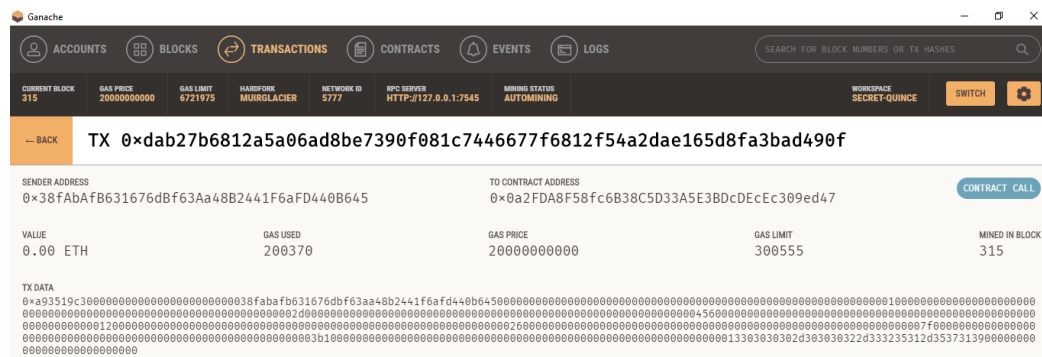


FIGURE 6.6: *Block generated for an entity registration*

blockchain, all nodes of the blockchain, record and verify the transaction's information (refer Fig. 6.7 and Fig. 6.8). The term "gas price" refers to the cost experienced by the sender for a specific transaction on the Ethereum blockchain in the proposed implementation. Miners adjust the price of gas based on the availability and demand of computing labour required to mine transactions. The miners are compensated with this gas price for validating transactions and adding blocks to the network. A higher price must be paid for a transaction that must be confirmed as soon as possible. The miners must be advised how much computational effort is required for a specific transaction. The gas limit, which specifies the maximum quantity of gas a sender can

FIGURE 6.7: *Transaction generated for a entity registration*

spend on a transaction, determines this. The gas cap also helps to keep fees paid to miners under control. Transactions may fail if the gas limit is set too low, because miners will ignore such transactions. Furthermore, the gas used in this transaction will be squandered. If the gas limit is exceeded and the transaction is completed at a lower gas price, the remaining gas will be returned to the sender's cryptocurrency wallet. Miners are important parts of the blockchain because they can either confirm a transaction or turn it down if the gas price is too low and doesn't meet their requirements. When a miner validates a transaction, the transaction is added to the miner's blockchain ledger. After other blockchain participants and miners have confirmed and validated the block added by the initial miner, a new block containing this transaction will be uploaded to the public blockchain. The transaction is known to be confirmed once it is published to a public blockchain. This confirmation procedure is similar to a voting mechanism in which additional miners and blockchain participants vote on the work of the first miners.

## 6.2.4 Fetching informetrics stored in blockchain

Here we have presented the detailed description of the DLT based system designed for fetching the unified informetrics for author, organization and journal.

### 6.2.4.1 Author level bibliometrics

Author information can be obtained in two ways after author credentials such as Orcid ID and bibliometrics have been deposited in the blockchain: (a) The publication

Transaction Details

Overview	Access List	State
[ This is a Rinkeby Testnet transaction only ]		
Transaction Hash:	0xc04748e64f8fb348611d3c50fcc173d6ca232c77d61318f4779b84d56a9239a	
Status:	Success	
Block:	9498764 34 Block Confirmations	
Timestamp:	8 mins ago (Oct-20-2021 02:44:19 PM +UTC)	
From:	0xde125abe485f76b16e2611bd79833a1d5eaa5a12	
To:	[Contract 0xd1caf38faa56720fbd8d13620f736eef56aebf9c Created]	
Value:	0 Ether (\$0.00)	
Transaction Fee:	0.00034672000312048 Ether (\$0.00)	
Gas Price:	0.000000001000000009 Ether (1.000000009 Gwei)	
Txn Type:	2 (EIP-1559)	
Gas Limit:	346,720	
Gas Used by Transaction:	346,720 (100%)	
Base Fee Per Gas:	9 wei (0.000000009 Gwei)	
Max Fee Per Gas:	0.000000001000000018 Ether (1.000000018 Gwei)	
Max Priority Fee Per Gas:	0.000000001 Ether (1 Gwei)	
Burnt Fees:	0.00000000000312048 Ether	

FIGURE 6.8: Number of confirmations received for a transaction generated for an entity registration

data of a certain author can be found using their Orcid ID. (b) The information of all the authors registered on the implemented informetrics system can also be retrieved. Fig. 6.9 (a) shows the admin dashboard, which lists the details of a certain author using the Orcid ID, and (b) displays all possible author credentials.

#### 6.2.4.2 Organization level bibliometrics

Organization information can be obtained in two ways after organization credentials such as organization name and bibliometrics have been deposited in the blockchain: (a) The publication data of a certain organization can be found by using their organization name. (b) The information of all the organizations registered on the implemented informetrics system can also be retrieved. Fig. 6.10 (a) shows the admin

**Fetch details with Orcid id**

Enter Orcid id:

Orcid ID	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
0000-0112-7858-3125	7	35	3	5	5	25

**Authors profiles registered on server**

Orcid ID	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
0000-0112-7858-3125	7	35	3	5	5	25
0000-0003-3793-4121	21	675	15	10	125	540
0000-0002-3251-5719	45	1110	18	38	127	945

FIGURE 6.9: Author retrieval for (a) particular author details (b) registered author's details.

dashboard, which lists the details of a certain organizations using the organization name, and (b) displays all possible organization credentials.

**Fetch details with Org. name**

Enter Org. name:

Org. Name	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
Indian Institute of Technology Madras	35449	591383	202	69278	208828	313277

**Org. profiles registered on server**

Org. Name	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
Indian Institute of Technology Madras	35449	591383	202	69278	208828	313277
Indian Institute of Science	54676	1236344	289	159065	455287	621992
Indian Institute of Technology Delhi	37873	747958	236	91757	296127	360074

FIGURE 6.10: Organization retrieval for (a) particular organization details (b) registered organization's details.



### 6.2.4.3 Journal level bibliometrics

Journal information can be obtained in two ways after journal credentials such as ISSN and bibliometrics have been deposited in the blockchain: (a) The publication data of a certain journal can be found by using their ISSN. (b) The information of all the journals registered on the implemented informetrics system can also be retrieved. Fig. 6.11 (a) shows the admin dashboard, which lists the details of a certain journal using the ISSN, and (b) displays all possible journal credentials.

**Fetch details with ISSN**

Enter ISSN:

ISSN	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
1063-6706	2804	228137	220	24208	133674	70255

**ISSN profiles registered on server**

ISSN	No. of papers	Total citations	<i>h</i> -index	Self citations	Repeated citations	Actual citations
1063-6706	2804	228137	220	24208	133674	70255
1361-8415	1689	98568	137	6058	45724	46786
2327-4662	2033	39331	80	5291	14386	19654

FIGURE 6.11: *Journal retrieval for (a) particular journal details (b) registered journal's details.*

## 6.3 Discussion and summary

In the first section, the analogy of applying distributed ledger technology in the publication industry is explained. The mapping of distributed ledger technology with key objects of the publication industry is introduced. Ledgers, Blocks, Records, and Transactions are mapped with key entities such as authors, organizations and journals for citations, research papers, and unique IDs. At the end of this section, a consensus mechanism named “Proof of bibliometric indicators” is introduced with its design and algorithm. In the last section, implementation details of distributed ledger technology

based on Truffle and Ganache are presented for author, organization and journal level bibliometrics.

In the next chapter, concluding remarks on the thesis are presented with its limitations and future scope.

---

---

# CHAPTER 7

---

## Conclusion and outlook

### 7.1 Concluding remarks

This thesis has investigated the three entities, mainly, author, organization, and journal, on the basis of data extracted from different indexing databases like Scopus and Web of Science. The question that this thesis is intended to answer is how one can combine the results of different indexing databases to generate unified informetric ledgers for three entities.

In this thesis, two indexing databases were utilized, Scopus and Web of Science, for extracting the article and citation information of authors, organizations, and journals. While initializing the work of citation analysis, there were different questions in our minds.

The first question was how to decide for which entity the data should be extracted, and the second question was where to extract the information about those entities. The third question was what should be the source of data for finalized entities. Should data from all indexing databases be extracted, or should a few of them be shortlisted first. The fourth question was that after extracting the data, what should

be kept and what should be ignored for the purpose of analysis and results. And the last question was how the analyzed data should be given the potential and features of distributed ledger technology.

Investigations were performed across various resources to start with the answer to the first question. A literature review was performed and it was found that the three prime entities which are always connected with the research publication industry were author, organization, and journal. Then, for the answer to the second question, various online resources were retrieved. There were two points, first identify the source and second decide the sample size. So, after completing the prime investigations, both points were answered and the complete derivation was performed.

For the third question, a number of indexing databases were accessed and explored. The websites and engines provided by various indexing databases to extract the required information from them were analyzed. The interests of different stakeholders were also studied in context with credibility and citation information available in indexing databases. Later on, literature was also studied, which was talking about different indexing databases for comparison, analysis, and result purposes. So this exercise and interest of different stakeholders helped us to decide that which two databases would be the part of our study.

The fourth question which was required to be answered was what to ignore and what to keep. During our insight study of data extracted from both indexing databases, it was observed that all data elements carried DOIs as a unique identifier element associated with them. So it was decided to keep everything that had a DOI number and ignore everything else. So, extraction of articles and citations was done only on the basis of DOI numbers.

After getting the answer for fourth question, last step was performed. The idea was to put final results into the distributed ledger so that required resultants would be empowered with distributed ledger technology.

A lot of literature surveys were done to write the answer to this question. The thought was to see what the actual potential of this technology is, and what other fields are taking advantage of this technology. After completing the literature review and applying the required mechanism, a novel approach to empower citation analysis with distributed ledger technology was identified.

Let's see the concluding investigations of the thesis, chapter by chapter, starting from Chapter 2 onwards.

Chapter 2 discussed the review of literature done for the identified problem. A review of literature is very important to understand the depth and significance of the identified problem. One can say that a review of literature also helps us to formulate the problem in a correct and rapid manner. Our problem has number of important elements like informetrics, indexing databases, distributed ledger technology and its applications. The literature review done for the informetrics talks about the different bibliometric indicators, their significance, their role, their credibility, and their importance. A literature review done for the different indexing databases helped us to analyse the different indexing databases in a detailed manner. Their size, importance, and role in calculating unified informetrics are also studied. Some computations were also performed for the calculation of the weight of citations. This was required because the outcomes of different indexing databases were merged. Then, a literature review based on distributed ledger technology is presented based on literature available on distributed ledger technology, distributed ledger technology-based applications, and the role of consensus mechanisms in distributed ledger technology. At the end of the chapter, a research gap was presented and the primary objectives based on which the identified problem will be solved were outlined.

Chapter 3 has dealt with the useful insights of  $h$ -index as an important informetrics.  $h$ -index focuses on both quantity and impact of publications but ignores highest citations received by the scientific work. In this chapter, a meaningful approach to study such limitations of  $h$ -index is followed. The study also presents and talks about other variants of  $h$ -index as well in the form of  $g$ -index. At the end, a new index or

a scientific approach named as  $h_c$  is proposed. This index takes care of limitations of  $h$ -index and provides a supplementary approach by adding weight of highest cited paper to  $h$ -index. The ranking of an individual within a discipline as well as in an organization can be considerably improved by using the weighted technique named as  $h_c$ , which can offer valuable insight to young or lower-ranked authors.

Chapter 4 is concerned with the linking of indexing databases and the generation of unified informetrics. This chapter gives a detailed view of the work done to solve the identified problem. In this chapter, it is specified how the uniqueness among different elements of entities is maintained. How an author as an entity, how an organization as an entity, and how a journal as an entity would be dealt with are described. This chapter also answers all the objectives defined during the formulation of our research work. ORCID as an author, ORG ID as an organization, and ISSN as a journal are used to remove any ambiguities in the research work. Mapping is a significant initiative in our research work. It gives us hope that a complete system can end up with a robust distributed ledger technology based application. Different components of distributed ledger technology are identified and mapped with the basic building blocks of research publication. Moving further with the idea, different sources used to collect the data for all three entities are presented. This is explained under the category of data description and filtering. In the next step, data extraction from different indexing databases is explained; how articles are extracted initially, how filtration is performed, and how the data set was finalized to perform extraction of citation information to move further with citation analysis. Complete methodology in the form of flowcharts and algorithms is presented. After completing the extraction of publication and citation information, unified informetrics such as  $h$ -index, impact factor, and citation count, etc., are calculated.

Chapter 5 dealt with the statistical analysis of Conflate. Results based on three entities are categorized. It was important to answer all the questions identified during the literature review with proper evidence. Existing data sets of Scopus and Web of Science on the basis of number of publications, then on the basis of number of

citations are compared. After comparing the number of publications and the number of citations, bibliometric indicators across Scopus, the Web of Science, and Conflate are calculated. In the final section, citations are compared using various parameters such as self, repeated, actual, and total to gain an understanding of the citation data of authors, organizations, and journals. Additionally, the study of the number of citations vs.  $h$ -index of an entity is presented. For journals, the impact factor is also analyzed.

Chapter 6 talked about the distributed ledger based implementation of Conflate. Initially, an idea of mapping distributed ledgers and the research publication industry is presented. A new consensus mechanism approach named “proof of bibliometric indicators” is explored in the form of a model and an algorithm. In the last section, the implementation details of an idea are presented with the support of the Ganache and Truffle frameworks. This implementation is performed on a local blockchain with Conflate data produced for various bibliometrics of an author.

## **7.2 How our work is overcoming the identified research gap during literature review?**

During the literature review, it was observed that distributed ledger technology has not been used to its full potential in the research publication industry. Neither any evidence nor methodology proposing the combination or merging of multiple indexing databases in any way was found. So a novel methodology was introduced where one can see the potential of distributed ledger technology in the research publication industry and citation analysis resulting in unified informetrics across multiple indexing databases is implemented.

Our work will address the concerns raised by various literature stating the facts that which indexing database is better, which indexing database should be selected for the studies related to the informetrics of individuals, organizations, journals, and countries as well, and which indexing database has the widest coverage.

The present work will also meet the data requirements raised by various accreditation agencies, funding bodies, ranking frameworks, and hiring agencies for individuals as well as groups.

Due to the generation of unified informetrics across multiple indexing databases, stakeholders may utilize the features of a unified database instead of taking and compiling the data from multiple indexing databases for their requirements.

At the end, the generated unified informetrics is deployed with the features of distributed ledger technology, resulting in the powerful perseverance and presentation of the generated data.

### 7.3 Summary of contribution

The contribution of the scientific work performed in the Ph.D. can be highlighted as follows:

- Introduction of  $h_c$  index, a supplementary approach to  $h$ -index for low ranked entities.
- Implementation of unified informetrics across multiple indexing databases, a one-stop solution for all stakeholders looking for hiring, promotion, funding, and ranking, etc.
- Presentation of consensus mechanism named “proof of bibliometric indicators”, to ensure the validity of records.
- Implementation of distributed ledger technology in the publication industry as an established mechanism to achieve the unified recording of informetrics.

### 7.4 Future direction

In this thesis, three entities and two indexing databases with the power of distributed ledger technology are used. The feasibility and effectiveness of each of the



presented approaches have been elaborated with detailed analysis.

However, there are still several possible areas for further exploration and extension. Here are some interesting areas for possible future developments and research.

- **Different indexing databases:** In this research, features of two indexing databases, such as Scopus and Web of Science, are studied. Hence, the performed study is limited to two indexing databases. One can extend the study further with the use of indexing databases like Microsoft Academics, Google Scholar, OpenAIRE, Mendeley, and Zenodo. According to the model, all of these indexing databases can be put together to get a single set of informetrics.
- **Different bibliometric indicators:** In our research, citation analysis on three entities (author, organization, and journal) is performed.  $h$ -index for all three entities and impact factor in the case of journals is calculated. Hence, the performed study is limited to only two bibliometric indicators. One can extend the study further with the calculation of different  $h$ -index variants on two or multiple indexing databases. In the case of journals, one can extend the study further with the calculation of an Eigenfactor score for journals of different disciplines.
- **Different technology aspects:** In this research, citation analysis is empowered with distributed ledger technology. Hence, the performed study is limited to the implementation of only one technology in research publication industry. One can extend the study further with the use of gamification and its gaming elements in the research publication industry. The use of gamification in the research publication industry can be helpful to increase the motivation and encouragement of its stakeholders for the extraction of unified informetrics for different entities.
- **Different ranking parameters:** In this research, a unified informetrics is calculated for all three entities based on different parameters. Although this work supports the publication, citation, and informetrics data required by different

ranking agencies, it does not provide any direct mechanism to rank any authors, organizations, or journals. Hence, the performed study is limited to providing the required credentials but can be further extended to propose a full fledged ranking mechanism for three or more entities.

---

## REFERENCES

- [1] Nisa Bakkalbasi, Kathleen Bauer, Janis Glover, and Lei Wang. Three options for citation tracking: Google scholar, scopus and web of science. *Biomedical digital libraries*, 3(1):7, 2006.
- [2] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic. *arXiv preprint arXiv:2005.10732*, 2020.
- [3] Gerson Pech and Catarina Delgado. Assessing the publication impact using citation data from both scopus and wos databases: an approach validated in 15 research fields. *Scientometrics*, 125(2):909–924, 2020.
- [4] Lutz Bornmann, K Brad Wray, and Robin Haunschild. Citation concept analysis (cca): a new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by thomas s. kuhn and karl r. popper. *Scientometrics*, 122(2):1051–1074, 2020.
- [5] Kate L Turabian. *A manual for writers of research papers, theses, and dissertations: Chicago style for students and researchers*. University of Chicago Press, 2013.
- [6] Charles Lipson. *Cite right: a quick guide to citation styles—MLA, APA, Chicago, the sciences, professions, and more*. University of Chicago Press, 2011.

- [7] Anita Papić. Informetrics: The development, conditions and perspectives. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 700–704. IEEE, 2017.
- [8] JEAN Tague-Sutcliffe. Quantitative methods in documentation. *Fifty years of information progress: a Journal of Documentation review*, pages 147–188, 1994.
- [9] William W Hood and Concepción S Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291, 2001.
- [10] Alberto Martín-Martín, Enrique Orduña-Malea, Juan M Ayllón, and Emilio Delgado Lopez-Cozar. The counting house: Measuring those who count. presence of bibliometrics, scientometrics, informetrics, webometrics and altmetrics in the google scholar citations, researcherid, researchgate, mendeley & twitter. *arXiv preprint arXiv:1602.02412*, 2016.
- [11] Siluo Yang, Qingli Yuan, Jiahui Dong, et al. Are scientometrics, informetrics, and bibliometrics different? *Data Science and Informetrics*, 1(01):50, 2020.
- [12] Daisy Jacobs. Demystification of bibliometrics, scientometrics, informetrics and webometrics. In *11th DIS Annual Conference*, volume 1, page 19, 2010.
- [13] MS Galyavieva. On the formation of the concept of informetrics. *Scientific and Technical Information Processing*, 40(2):89–96, 2013.
- [14] Times Higher Education Staff. Times higher education announces reforms to its world university rankings, 2014. URL <https://www.timeshighereducation.com/news/times-higher-education-announces-reforms-to-its-world-university-rankings/2017071.article>.
- [15] Kiran Sharma and Parul Khurana. Growth and dynamics of econophysics: a bibliometric and network analysis. *Scientometrics*, 126(5):4417–4436, 2021.
- [16] Lokman I Meho and Cassidy R Sugimoto. Assessing the scholarly impact of information studies: A tale of two citation databases—scopus and web of science.

- Journal of the American Society for information science and technology*, 60(12):2499–2508, 2009.
- [17] Éric Archambault, David Campbell, Yves Gingras, and Vincent Larivière. Comparing bibliometric statistics obtained from the web of science and scopus. *Journal of the American society for information science and technology*, 60(7):1320–1326, 2009.
- [18] Ben Martin. The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3):343–362, 1996.
- [19] Judit Bar-Ilan, Mark Levene, and Ayelet Lin. Some measures for comparing citation databases. *Journal of Informetrics*, 1(1):26–34, 2007.
- [20] Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, et al. A data citation roadmap for scientific publishers. *Scientific data*, 5(1):1–11, 2018.
- [21] Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386, 2020.
- [22] Caroline Birkle, David A Pendlebury, Joshua Schnell, and Jonathan Adams. Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1):363–376, 2020.
- [23] Richard Van Noorden. Google scholar pioneer on search engine’s future. *Nature*, 10, 2014.
- [24] Michael Gusenbauer. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214, 2019.

- 
- [25] Emilio Delgado López-Cózar, Enrique Orduña-Malea, and Alberto Martín-Martín. Google scholar as a data source for research assessment. In *Springer handbook of science and technology indicators*, pages 95–127. Springer, 2019.
- [26] Christoph Neuhaus and Hans-Dieter Daniel. Data sources for performing citation analysis: an overview. *Journal of documentation*, 2008.
- [27] Raminta Pranckutė. Web of science (wos) and scopus: The titans of bibliographic information in today’s academic world. *Publications*, 9(1):12, 2021.
- [28] Sandra L De Groote and Rebecca Raszewski. Coverage of google scholar, scopus, and web of science: A case study of the h-index in nursing. *Nursing outlook*, 60(6):391–400, 2012.
- [29] Dag W Aksnes and Gunnar Sivertsen. A criteria-based assessment of the coverage of scopus and web of science. *Journal of Data and Information Science*, 4(1):1–21, 2019.
- [30] Mike Thelwall. Dimensions: A competitor to scopus and the web of science? *Journal of informetrics*, 12(2):430–435, 2018.
- [31] Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado-López-Cózar. Google scholar, web of science, and scopus: which is best for me? *Impact of Social Sciences Blog*, 2019.
- [32] Lutishoor Salisbury. Web of science and scopus: A comparative review of content and searching capabilities. *The Charleston Advisor*, 11(1):5–18, 2009.
- [33] Tomaz Bartol, Gordana Budimir, Doris Dekleva-Smrekar, Miro Pusnik, and Primoz Juznic. Assessment of research fields in scopus and web of science in the view of national research evaluation in slovenia. *Scientometrics*, 98(2):1491–1504, 2014.
- [34] Fiorenzo Franceschini, Domenico Maisano, and Luca Mastrogiacomo. Empirical analysis and classification of database errors in scopus and web of science. *Journal of informetrics*, 10(4):933–953, 2016.

- 
- [35] Weishu Liu, Meiting Huang, and Haifeng Wang. Same journal but different numbers of published records indexed in scopus and web of science core collection: causes, consequences, and solutions. *Scientometrics*, 126(5):4541–4550, 2021.
- [36] Kiduk Yang and Lokman I Meho. Citation analysis: a comparison of google scholar, scopus, and web of science. *Proceedings of the American Society for information science and technology*, 43(1):1–15, 2006.
- [37] Henk F Moed, Valentina Markusova, and Mark Akoev. Trends in russian research output indexed in scopus and web of science. *Scientometrics*, 116(2):1153–1180, 2018.
- [38] Lokman I Meho and Yvonne Rogers. Citation counting, citation ranking, and h-index of human-computer interaction researchers: a comparison of scopus and web of science. *Journal of the American Society for Information Science and Technology*, 59(11):1711–1726, 2008.
- [39] Carmen López-Illescas, Félix de Moya Anegón, and Henk F Moed. Comparing bibliometric country-by-country rankings derived from the web of science and scopus: The effect of poorly cited journals in oncology. *Journal of information science*, 35(2):244–256, 2009.
- [40] MGCSA Walport et al. Distributed ledger technology: Beyond blockchain. *UK Government Office for Science*, 1:1–88, 2016.
- [41] Roger Maull, Phil Godsiff, Catherine Mulligan, Alan Brown, and Beth Kewell. Distributed ledger technology: Applications and implications. *Strategic Change*, 26(5):481–489, 2017.
- [42] Nabil El Ioini and Claus Pahl. A review of distributed ledger technologies. In *OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”*, pages 277–288. Springer, 2018.

- 
- [43] Arif Perdana, Alastair Robb, Vivek Balachandran, and Fiona Rohde. Distributed ledger technology: Its evolutionary path and the road ahead. *Information & Management*, page 103316, 2020.
- [44] JP Puntinx. Distributed ledger technology vs blockchain technology. *Viitattu*, 1:2017, 2017.
- [45] Dimitris Chatzopoulos, Anurag Jain, Sujit Gujar, Boi Faltings, and Pan Hui. Towards mobile distributed ledgers. *arXiv preprint arXiv:2101.04825*, 2021.
- [46] Kiran Sharma and Parul Khurana. Emerging trends and collaboration patterns unveil the scientific production in blockchain technology: A bibliometric and network analysis from 2014-2020. *arXiv preprint arXiv:2110.01871*, 2021.
- [47] Parul Khurana, Geetha Ganesan, Gulshan Kumar, and Kiran Sharma. Exploring the potential of distributed ledger technology in publication industry – a technological review. In *CEUR Workshop Proc. 2869*, pages 32–40, 2021.
- [48] Seyed Mojtaba Hosseini Bamakan, Amirhossein Motavali, and Alireza Babaei Bondarti. A survey of blockchain consensus algorithms performance evaluation criteria. *Expert Systems with Applications*, page 113385, 2020.
- [49] Natalia Chaudhry and Muhammad Murtaza Yousaf. Consensus algorithms in blockchain: comparative analysis, challenges and opportunities. In *2018 12th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 54–63. IEEE, 2018.
- [50] Christian Cachin and Marko Vukolić. Blockchain consensus protocols in the wild. *arXiv preprint arXiv:1707.01873*, 2017.
- [51] Parul Khurana, Kiran Sharma, and Kiran Khatter. Proof of bibliometric indicators: a blockchain based consensus protocol for publications. *Multimedia Tools and Applications*, pages 1–16, 2022.
- [52] Lakshmi Siva Sankar, M Sindhu, and M Sethumadhavan. Survey of consensus protocols on blockchain applications. In *2017 4th International Conference on*



- Advanced Computing and Communication Systems (ICACCS)*, pages 1–5. IEEE, 2017.
- [53] Arshdeep Singh, Gulshan Kumar, Rahul Saha, Mauro Conti, Mamoun Alazab, and Reji Thomas. A survey and taxonomy of consensus protocols for blockchains. *Journal of Systems Architecture*, page 102503, 2022.
- [54] Peter Jacso. As we may search—comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases. *Current science*, 89(9):1537–1547, 2005.
- [55] Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.
- [56] Judit Bar-Ilan. Which h-index?—a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.
- [57] Péter Jacsó. The h-index for countries in web of science and scopus. *Online information review*, 2009.
- [58] Elizabeth Vieira and José Gomes. A comparison of scopus and web of science for a typical university. *Scientometrics*, 81(2):587–600, 2009.
- [59] Judit Bar-Ilan. Citations to the “introduction to informetrics” indexed by wos, scopus and google scholar. *Scientometrics*, 82(3):495–506, 2010.
- [60] Leslie S Adriaanse and Chris Rensleigh. Web of science, scopus and google scholar. *The Electronic Library*, 2013.
- [61] Philippe Mongeon and Adèle Paul-Hus. The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228, 2016.
- [62] Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016.

- [63] Juan Gorraiz, David Melero-Fuentes, Christian Gumpenberger, and Juan-Carlos Valderrama-Zurián. Availability of digital object identifiers (dois) in web of science and scopus. *Journal of informetrics*, 10(1):98–109, 2016.
- [64] Aparna Basu, Sumit Kumar Banshal, Khushboo Singhal, and Vivek Kumar Singh. Designing a composite index for research performance evaluation at the national or regional level: ranking central universities in india. *Scientometrics*, 107(3):1171–1193, 2016.
- [65] Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. *Journal of informetrics*, 12(4):1160–1177, 2018.
- [66] Alberto Martín-Martín, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. Coverage of highly-cited documents in google scholar, web of science, and scopus: a multidisciplinary comparison. *Scientometrics*, 116(3):2175–2188, 2018.
- [67] Saif Aldeen S AlRyalat, Lna W Malkawi, and Shaher M Momani. Comparing bibliometric analysis using pubmed, scopus, and web of science databases. *JoVE (Journal of Visualized Experiments)*, (152):e58494, 2019.
- [68] Miguel-Angel Vera-Baceta, Michael Thelwall, and Kayvan Kousha. Web of science and scopus language coverage. *Scientometrics*, 121(3):1803–1813, 2019.
- [69] Enrique Orduna-Malea, Selenay Aytac, and Clara Y Tran. Universities through the eyes of bibliographic databases: a retroactive growth comparison of google scholar, scopus and web of science. *Scientometrics*, 121(1):433–450, 2019.
- [70] Junwen Zhu and Weishu Liu. A tale of two databases: the use of web of science and scopus in academic papers. *Scientometrics*, pages 1–15, 2020.
- [71] Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. Google scholar, microsoft academic, scopus, dimensions, web

- of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906, 2021.
- [72] Vivek Kumar Singh, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126(6):5113–5142, 2021.
- [73] Parul Khurana and Kiran Sharma. Impact of h-index on author's rankings: an improvement to the h-index for lower-ranked authors. *Scientometrics*, pages 1–16, 2022.
- [74] Eugene Garfield. Citation indexes for science. a new dimension in documentation through association of ideas. *International journal of epidemiology*, 35(5):1123–1127, 2006.
- [75] Blaise Cronin, Herbert Snyder, and Helen Atkins. Comparative citation rankings of authors in monographic and journal literature: A study of sociology. *Journal of documentation*, 1997.
- [76] Lutz Bornmann and Hans-Dieter Daniel. Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392, 2005.
- [77] Jean-Francois Molinari and Alain Molinari. A new methodology for ranking scientific institutions. *Scientometrics*, 75(1):163–174, 2008.
- [78] Lutz Bornmann. Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, 73(5):775–787, 2017.
- [79] Philip Ball. Index aims for fair ranking of scientists, 2005.
- [80] Belle Dumé. How high is your h-index? *Physics World*, 18(9):7, 2005.
- [81] Wolfgang Glänzel. On the h-index—a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2):315–321, 2006.

- 
- [82] Majdi Maabreh and Izzat M Alsmadi. A survey of impact and citation indices: Limitations and issues. *International Journal of Advanced Science and Technology*, 40(4):35–54, 2012.
- [83] Tibor Braun, Wolfgang Glänzel, and Andras Schubert. A hirsch-type index for journals. *The scientist*, 19(22):8, 2005.
- [84] Tünde Gracza and Istvánné Somoskövi. Impact factor and/or hirsch index? *Orvosi Hetilap*, 148(18):849–852, 2007.
- [85] Tibor Braun, Wolfgang Glänzel, and András Schubert. A hirsch-type index for journals. *Scientometrics*, 69(1):169–173, 2006.
- [86] Leo Egghe. The influence of transformations on the h-index and the g-index. *Journal of the American Society for Information Science and Technology*, 59(8):1304–1312, 2008.
- [87] Giovanni Abramo, Ciriaco Andrea D’Angelo, and Fulvio Viel. A robust benchmark for the h-and g-indexes. *Journal of the American Society for Information Science and Technology*, 61(6):1275–1280, 2010.
- [88] Ludo Waltman. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391, 2016.
- [89] Jorge E Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.
- [90] Marcel Dunaiski, Jaco Geldenhuys, and Willem Visser. Globalised vs averaged: Bias and ranking performance on the author level. *Journal of Informetrics*, 13(1):299–313, 2019.
- [91] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

- 
- [92] Michal Nykl, Michal Campr, and Karel Ježek. Author ranking based on personalized pagerank. *Journal of Informetrics*, 9(4):777–799, 2015.
- [93] Marcel Dunaiski, Willem Visser, and Jaco Geldenhuys. Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, 10(2):392–407, 2016.
- [94] Marcel Dunaiski, Jaco Geldenhuys, and Willem Visser. Author ranking evaluation at scale. *Journal of Informetrics*, 12(3):679–702, 2018.
- [95] Sergio Alonso, Francisco Javier Cabrerizo, Enrique Herrera-Viedma, and Francisco Herrera. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics*, 82(2):391–400, 2010.
- [96] Bihui Jin, Liming Liang, Ronald Rousseau, and Leo Egghe. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, 52(6):855–863, 2007.
- [97] Chun-Ting Zhang. The e-index, complementing the h-index for excess citations. *PLoS One*, 4(5):e5429, 2009.
- [98] Chun-Ting Zhang. The h'-index, effectively improving the h-index based on the citation distribution. *PloS one*, 8(4):e59912, 2013.
- [99] Chun-Ting Zhang. A proposal for calculating weighted citations based on author rank. *EMBO reports*, 10(5):416–417, 2009.
- [100] Muhammad Usman, Ghulam Mustafa, and Muhammad Tanvir Afzal. Ranking of author assessment parameters using logistic regression. *Scientometrics*, pages 1–19, 2020.
- [101] Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics*, 1(3):193–203, 2007.
- [102] Pablo D Batista, Mônica G Campiteli, and Osame Kinouchi. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1): 179–189, 2006.

- 
- [103] Peter Vinkler. Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of information science*, 33(4):481–491, 2007.
- [104] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter 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. *Journal of the American society for information science and technology*, 59(5):830–837, 2008.
- [105] Michael Schreiber. An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the a-index, and the r-index. *Journal of the American Society for Information Science and Technology*, 59(9):1513–1522, 2008.
- [106] Rodrigo Costas and María Bordons. Is g-index better than h-index? an exploratory study at the individual level. *Scientometrics*, 77(2):267–288, 2008.
- [107] Leo Egghe. An improvement of the h-index: The g-index. *ISSI newsletter*, 2(1):8–9, 2006.
- [108] Michael Schreiber. The influence of self-citation corrections on egghe’sg index. *Scientometrics*, 76(1):187–200, 2008.
- [109] Richard SJ Tol. A rational, successive g-index applied to economics departments in ireland. *Journal of Informetrics*, 2(2):149–155, 2008.
- [110] Jingda Ding, Chao Liu, and Goodluck Asobenie Kandonga. Exploring the limitations of the h-index and h-type indexes in measuring the research performance of authors. *Scientometrics*, 122(3):1303–1322, 2020.
- [111] Leo Egghe. The hirsch index and related impact measures. *Annual review of information science and technology*, 44(1):65–114, 2010.
- [112] Michael Schreiber. Twenty hirsch index variants and other indicators giving more or less preference to highly cited papers. *Annalen der Physik*, 522(8):536–554, 2010.

- 
- [113] Hadi Esmaily, Elmira Niknami, and Ali Saffaei. Comment on: H-index is an ugly truth; but what about other scientometric criteria? *Shiraz E-Medical Journal*, (In Press), 2019.
- [114] Massimo Franceschet. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10):1992–2012, 2011.
- [115] Deepti Gupta and Navneet Gupta. Higher education in india: structure, statistics and challenges. *Journal of education and Practice*, 3(2), 2012.
- [116] Massimo Franceschet. A comparison of bibliometric indicators for computer science scholars and journals on web of science and google scholar. *Scientometrics*, 83(1):243–258, 2010.
- [117] Nees Jan van Eck and Ludo Waltman. Accuracy of citation data in web of science and scopus. *arXiv preprint arXiv:1906.07011*, 2019.
- [118] Rajesh Chandrakar. Digital object identifier system: an overview. *The Electronic Library*, 2006.
- [119] Rogério Mugnaini, Grischa Fraumann, Esteban F Tuesta, and Abel L Packer. Openness trends in brazilian citation data: factors related to the use of dois. *Scientometrics*, 126(3):2523–2556, 2021.
- [120] Nicholas Homenda. Persistent urls and citations offered for digital objects by digital libraries. *Information Technology and Libraries*, 40(2), 2021.
- [121] Erin Carreiro. Electronic books: how digital devices and supplementary new technologies are changing the face of the publishing industry. *Publishing research quarterly*, 26(4):219–235, 2010.
- [122] Stephen Mooney. Digital object identifiers for ebooks: What are we identifying? *Publishing research quarterly*, 17(1):29–36, 2001.

- [123] Parul Khurana, Geetha Ganesan, Gulshan Kumar, and Kiran Sharma. A bibliometric analysis to unveil the impact of digital object identifiers (doi) on bibliometric indicators. In Pradeep Kumar Singh, Sławomir T. Wierzchoń, Sudeep Tanwar, Joel J. P. C. Rodrigues, and Maria Ganzha, editors, *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security*, pages 859–869, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-19-1142-2.
- [124] Ashok Agarwal, Damayanthi Durairajanayagam, Sindhuja Tatagari, Sandro C Esteves, Avi Harlev, Ralf Henkel, Shubhadeep Roychoudhury, Sheryl Homa, Nicolás Garrido Puchalt, Ranjith Ramasamy, et al. Bibliometrics: tracking research impact by selecting the appropriate metrics. *Asian journal of andrology*, 18(2):296, 2016.
- [125] Ernesto Roldan-Valadez, Shirley Yoselin Salazar-Ruiz, Rafael Ibarra-Contreras, and Camilo Rios. Current concepts on bibliometrics: a brief review about impact factor, eigenfactor score, citescore, scimago journal rank, source-normalised impact per paper, h-index, and alternative metrics. *Irish Journal of Medical Science (1971-)*, pages 1–13, 2019.
- [126] James Fowler and Dag Aksnes. Does self-citation pay? *Scientometrics*, 72(3): 427–437, 2007.
- [127] Michael Norris and Charles Oppenheim. The h-index: A broad review of a new bibliometric indicator. *Journal of Documentation*, 2010.
- [128] Martin Szomszor, David A Pendlebury, and Jonathan Adams. How much is too much? the difference between research influence and self-citation excess. *Scientometrics*, 123(2):1119–1147, 2020.
- [129] Valeri Craigle. Adopting doi in legal citation: A roadmap for the legal academy. *Legal Citation: A Roadmap for the Legal Academy (March 8, 2021)*. *Legal Reference Services Quarterly (2021 Forthcoming)*, University of Utah College of Law Research Paper Forthcoming, 2021.



- [130] Michael E Rose and John R Kitchin. pybliometrics: Scriptable bibliometrics using a python interface to scopus. *SoftwareX*, 10:100263, 2019.
- [131] Michael E. Rose. Compute scholarly metrics in python with pandas and numpy edit, 2017. URL <https://pypi.org/project/scholarmetrics/>.
- [132] Enrico Bacis. enricobacis/wos, 2019. URL <https://github.com/enricobacis/wos>.
- [133] Christoph Bartneck and Servaas Kokkelmans. Detecting h-index manipulation through self-citation analysis. *Scientometrics*, 87(1):85–98, 2011.
- [134] Hui Li and Weishu Liu. Same same but different: self-citations identified through scopus and web of science core collection. *Scientometrics*, 124(3):2723–2732, 2020.
- [135] Budiman Minasny, Alfred E Hartemink, Alex McBratney, and Ho-Jun Jang. Citations and the h index of soil researchers and journals in the web of science, scopus, and google scholar. *PeerJ*, 1:e183, 2013.
- [136] Ted Brown and Sharon A Gutman. A comparison of bibliometric indicators in occupational therapy journals published in english. *Canadian Journal of Occupational Therapy*, 86(2):125–135, 2019.
- [137] Garey A Fox, Amanda K Fox, and Lucie Guertault. A case study on the relevance of the journal impact factor. *Transactions of the ASABE*, 63(2):243–249, 2020.
- [138] Mario Cantín, M Muñoz, and Ignacio Roa. Comparison between impact factor, eigenfactor score, and scimago journal rank indicator in anatomy and morphology journals. *International Journal of Morphology*, 33(3), 2015.
- [139] Parul Khurana, Geetha Ganesan, Gulshan Kumar, and Kiran Sharma. A comparative analysis of unified informetrics with scopus and web of science. *Journal of Scientometric Research*, 11(2):146–154, 2022.

- 
- [140] Ghassan O Karame and Elli Androulaki. *Bitcoin and blockchain security*. Artech House, 2016.
- [141] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260, 2008.
- [142] Gareth RT White. Future applications of blockchain in business and management: A delphi study. *Strategic Change*, 26(5):439–451, 2017.
- [143] Christoph Prybila, Stefan Schulte, Christoph Hochreiner, and Ingo Weber. Runtime verification for business processes utilizing the bitcoin blockchain. *Future Generation Computer Systems*, 107:816–831, 2020.
- [144] NC Rainer Böhme. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2), 2015.
- [145] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman, et al. Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2(6-10):71, 2016.
- [146] Svein Ølnes, Jolien Ubacht, and Marijn Janssen. Blockchain in government: Benefits and implications of distributed ledger technology for information sharing, 2017.
- [147] Jesse Yli-Huumo, Deokyoon Ko, Sujin Choi, Sooyong Park, and Kari Smolander. Where is current research on blockchain technology?—a systematic review. *PloS one*, 11(10):e0163477, 2016.
- [148] Chris Khan, Antony Lewis, Emily Rutland, Clemens Wan, Kevin Rutter, and Clark Thompson. A distributed-ledger consortium model for collaborative innovation. *Computer*, 50(9):29–37, 2017.
- [149] Iman Musleh, Samer Zain, Mamoun Nawahdah, and Norsaremah Salleh. Automatic generation of android sqlite database components. In *SoMeT*, pages 3–16, 2018.

- [150] Shuai Zeng, Xiaochun Ni, Yong Yuan, and Fei-Yue Wang. A bibliometric analysis of blockchain research. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 102–107. IEEE, 2018.
- [151] Richard Adams, Beth Kewell, and Glenn Parry. Blockchain for good? digital ledger technology and sustainable development goals. In *Handbook of sustainability and social science research*, pages 127–140. Springer, 2018.
- [152] Mohammad Dabbagh, Mehdi Sookhak, and Nader Sohrabi Safa. The evolution of blockchain: A bibliometric study. *Ieee Access*, 7:19212–19221, 2019.
- [153] Haydar Yalcin and Tugrul Daim. Mining research and invention activity for innovation trends: case of blockchain technology. *Scientometrics*, 126(5):3775–3806, 2021.
- [154] Yi-Ming Guo, Zhen-Ling Huang, Ji Guo, Xing-Rong Guo, Hua Li, Meng-Yu Liu, Safa Ezzeddine, and Mpeoane Judith Nkeli. A bibliometric analysis and visualization of blockchain. *Future Generation Computer Systems*, 116:316–332, 2021.
- [155] Anushree Tandon, Puneet Kaur, Matti Mäntymäki, and Amandeep Dhir. Blockchain applications in management: A bibliometric analysis and literature review. *Technological Forecasting and Social Change*, 166:120649, 2021.
- [156] Alexis Collomb and Klara Sok. Blockchain/distributed ledger technology (dlt): What impact on the financial sector? *Digiworld Economic Journal*, (103), 2016.
- [157] John Naughton. Is blockchain the most important it invention of our age. *The Guardian*, 24, 2016.
- [158] Arif Perdana, Alastair Robb, Vivek Balachandran, and Fiona Rohde. Distributed ledger technology: Its evolutionary path and the road ahead. *Information & Management*, 58(3):103316, 2021.

---

# Publications and Patents

## Published/Accepted articles

[1] Parul Khurana, Geetha Ganesan, Gulshan Kumar, Kiran Sharma (2022). A comparative analysis of unified informetrics with Scopus and Web of Science. *Scientometric Research*. (Published, Scopus indexed , Impact Factor 1.11, 2022)

[2] Parul Khurana, Gulshan Kumar, Geetha Ganesan (2021). Exploring the potential of Distributed Ledger Technology in publication industry – A technological review. In: *CEUR workshop proceedings*, vol 2869, pp 32–40. ISSN 1613-0073. (Published, Scopus indexed)

[3] Parul Khurana, Geetha Ganesan, Gulshan Kumar, Kiran Sharma (2023). A Bibliometric Analysis to Unveil the Impact of Digital Object Identifiers (DOI) on Bibliometric Indicators. In: *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security. Lecture Notes in Networks and Systems*, vol 421. Springer, Singapore. (Published, Scopus indexed)

[4] Parul Khurana, Kiran Sharma (2022). Impact of h-index on author's rankings: An improvement to the h-index for lower-ranked authors. *Scientometrics*. (Published, SCI indexed, Impact Factor 3.801, 2022).

[5] Parul Khurana, Kiran Sharma, Kiran Khatter (2022). Proof of Bibliometric Indicators: A blockchain based consensus protocol for publications. *Multimedia Tools and Applications*. (Published, SCI Indexed, Impact Factor 2.577, 2022).

[6] Kiran Sharma, Parul Khurana (2021). Growth and dynamics of Econophysics: a bibliometric and network analysis. *Scientometrics*, 126(5), 4417-4436. *Scientometrics*. (Published, SCI Indexed, Impact Factor 3.801, 2021).

[7] Parul Khurana, Kiran Sharma (2023). Exploring the role of AI, IoT and BC during Covid-19: A bibliometric and network analysis. (Accepted, Book chapter under publication, Scopus Indexed, 2023).

### **Communicated articles**

[1] Parul Khurana, Kiran Sharma (2023). Emerging trends and collaboration patterns unveil the scientific production in blockchain technology: A bibliometric and network analysis from 2014-2020. (Communicated, 2023).

[2] Kiran Sharma, Parul Khurana, Gulshan Kumar (2023). Female authorship in Covid-19 retracted publications. (Communicated, 2023).

### **Patents**

[1] Patent published in patent journal, *Methods and Systems for Maintaining Records Associated with An Entity* (202011040067). (2020)