iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54661

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Publications Information Retrieval and Management System

T.Aparna[1], Shivani D[2], Pratiksha B[3], Nuzzath Tahreen[4], Anusha T[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Student, Department of Information Technology, G Narayanamma Institute of Technology and Science, Hyderabad, India.*

Abstract: ***In a world where technology is advancing rapidly, it is a difficult task to manually maintain research profiles in any educational institution. Faculty have to go through an exhaustive ordeal to keep up to pace with the ever-changing demands of the publication committees. Hence, there is a significant need for an interface allowing researchers to easily sort, maintain and download their list of publications with minimum work from the user end. This paper attempts to set up an interface that would use web scraping with selenium and python modules to link a researcher's list of publications present on Google Scholar to a PostgreSQL database and Excel application, allowing them to access and manipulate their research profiles in minimal steps.***
Index Terms: Web Scraping, Selenium, Python, Django, PostgreSQL, Google Scholar, Scopus, WebOfScience.

I. INTRODUCTION

Academic Institutions often face several problems in maintaining their research profiles. In the current scenario, a person has to manually copy and paste the publication details from Google Scholar to Excel sheets or Word documents according to their preferred presentation medium. It often involves the laborious task of manually extracting the papers they have written along with the number of citations, year of publication, publication type, publisher name, and journal/conference/book name. It is coupled with the drudging task of placing this data individually into an excel sheet. In the last year, millions of scientific research papers were published. Considering this massive scale and the underlying problems faced by each academic institution, there is a need for a user-friendly interface that dynamically extracts the publication details from Google Scholar through web scraping and allows the user to view, update and download their data into an excel sheet in a minimum number of steps.

II. LITERATURE SURVEY

The interface is developed entirely in Python using Django Framework. The paper [5] UNO: A Web Application using Django helped in learning the process of creating the interface. It gave complete insight into creating a project using the Django framework. [6] Performance Analysis of PostgreSQL, MySQL, Microsoft SQL Server Systems Based on TPC-H Tests gave analysis on different types of databases that can be used and helped in choosing PostgreSQL which can store huge data and has many features like data integrity, fault-tolerant environment. [1] Web Scraping and Data Acquisition Using Google Scholar gave insights on scraping data from Google Scholar. This paper focused on publication data acquisition from Google Scholar to database and excel sheet using Web Scraping with Beautiful Soup. The details of journals and conferences like impact factor, h-index which are not available on Google Scholar website cannot be extracted. The above paper is not providing searching options for users to extract publications based on faculty name, ISBN number, Publication title etc. The proper scraping method should be needed which allows movement from one page to another page and scrape multiple pages at a time. Scraping using selenium is chosen after analyzing the paper [4] Web Scraping based Product Comparison Model for E-Commerce Websites.

III. PROPOSED SYSTEM

A. *Objectives*

This paper aims to provide an approach for web scraping concerning Google Scholar, Scopus, and WebOfScience and use the data acquired to build a user-friendly interface allowing users to easily access their publication details.
The above-mentioned aims produce the following objectives:

1) To analyze and extract key HTML links from Google Scholar, Scopus, and WebOfScience.
2) To extract the publication details as per institution requirements.
3) To create a database on PostgreSQL and store the extracted information.
4) To check whether the Published paper is Scopus and WebOfscience indexed.

5) To retrieve and display the stored information to the user with comfortable choices.

The long-term goal of this paper is to provide the institutions with a user-friendly interface that would minimize the time it would take them to complete these tasks.

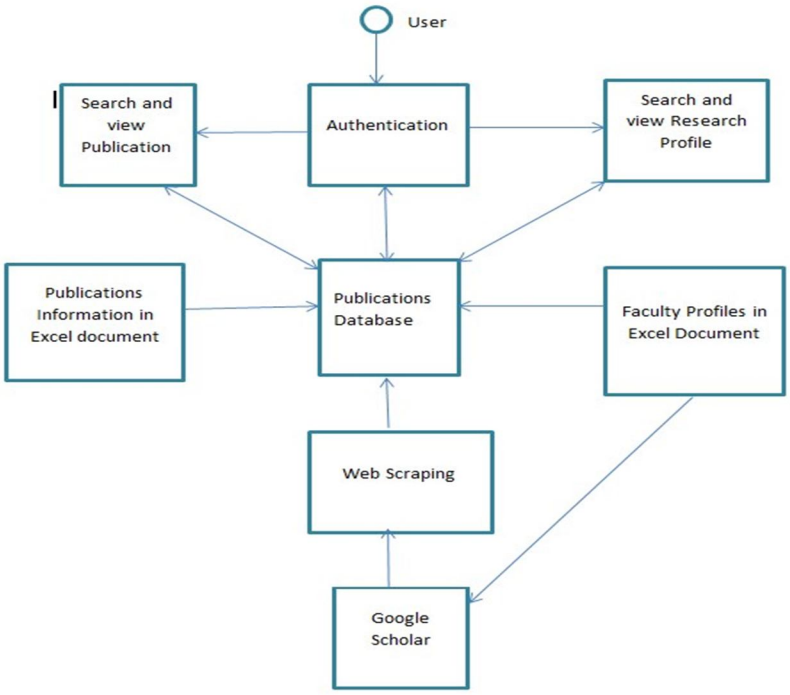


Fig-1 System Architecture

B. Methodology And Implementation

The following methodology was adopted in designing this interface.

1) Firstly, we need to store the URLs of different authors in an external file like an excel sheet, so that the URL of the page has to be accessed to fetch the details from Google Scholar.

2) The Google Scholar page which gives the list of articles that has been written by a particular author has basic HTML architecture. Web Scraping is the process of extracting information from web pages. Python provides simple methods to scrape any page.

The following interface was created in the following fragments:

a) The interface was created using advanced python (Django framework) and PostgreSQL database.

b) The first step is to have python installed in the system and it should have the following python modules installed in it – Selenium, CSV, requests, and Django. If not preinstalled user can install it using pip install command. Install the latest version of PostgreSQL. To use the PostgreSQL database update the settings.py file.

```
settings.py > ...

DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.postgresql_psycopg2',
        'NAME': 'MP_db',
        'USER': 'postgres',
        'PASSWORD': '12345',
        'HOST': '127.0.0.1',
        'PORT': '5432',
    }
}
```

Fig 2 PostgreSQL database

c) The next step is to create an excel sheet that contains the faculty ids and their respective Google scholar URLs in an indexed manner which can be further used to fetch the URL.

d) Then create a user-friendly interface with HTML, CSS, JavaScript, and Bootstrap.

e) Create a table consisting of attributes with corresponding data types to store scraped publication details in the PostgreSQL database.

f) The next part is to connect to the database by registering the table in the models.py file.

```
models.py > ...
class login_db(models.Model):
    username = models.CharField(max_length=15, primary_key=True)
    password = models.CharField(max_length=8)
    def __str__(self):
        return str(self.username)

class author_ids(models.Model):
    faculty_id = models.ForeignKey(login_db, on_delete=models.CASCADE)
    author_id = models.CharField(max_length=25, null=True, unique=True)

class publication(models.Model):
    faculty_id = models.ForeignKey(login_db, on_delete=models.CASCADE)
    faculty_name = models.CharField(max_length=50)
    title = models.CharField(max_length=100)
    authors = models.CharField(max_length=200, default="-")
    publication_type = models.CharField(max_length=30, default="-")
    name = models.CharField(max_length=150, default="-")
    publisher_name = models.CharField(max_length=100, default="-")
    published_yr = models.CharField(max_length=4, default="-")
    issn_no = models.CharField(max_length=15, default="--")
    impact_factor = models.FloatField(default=0)
    h_index = models.IntegerField(default=0)
    citations = models.IntegerField(default=0)
    Scopus = models.IntegerField(default=0)
    WebOfScience = models.IntegerField(default=0)
```

Fig 3 Publications table creation in Models.py

g) Using the selenium module, a chrome web driver is created to scrape the specific URL and simple methods can be used to extract the required details.

h) Then the PostgreSQL commands have to be written to insert the extracted details.

i) Finally check the output compatibility with the database table. This completes the creation of the interface.

K] bigint	faculty_id_id character varying (15)	faculty_name character varying (50)	title character varying (200)
1	FA1207	sesha bhargavi Velagaleti	A hybrid secure routing scheme for MANETS
2	FA1207	sesha bhargavi Velagaleti	A trust based secure routing scheme for MANETS
3	FA1207	sesha bhargavi Velagaleti	Detection of Multiple Malicious Nodes in MANETS in a Single Query
4	FA1207	sesha bhargavi Velagaleti	Enhance safety and security system for children in school campus by using wearable senso
5	FA1207	sesha bhargavi Velagaleti	Challenges in handling imbalanced big data: a survey
6	FA1207	sesha bhargavi Velagaleti	A Simulation And Analysis Of Secured S-DSR Protocol In Mobile Ad Hoc Networks
7	FA1207	sesha bhargavi Velagaleti	A novel method for trust evaluation in a mobile Ad Hoc network
8	FA1207	sesha bhargavi Velagaleti	A Simulation and Analysis Of Secured AODV Protocol in Mobile Ad Hoc Networks
9	FA1207	sesha bhargavi Velagaleti	Recommendation Based P2P File Sharing on Disconnected MANET
10	FA1207	sesha bhargavi Velagaleti	unreliable image material screening with contrast enhancement
11	FA1207	sesha bhargavi Velagaleti	Privacy Preserving and fully Anonymous Protocols for Profile matching in Mobile Social Net
12	FA1207	sesha bhargavi Velagaleti	A simulation and analysis of secured DSR protocol in mobile Ad hoc networks

rows: 124 of 124 Query complete 00:00:00.402 Ln 1, Col 1

Fig 4 Acquisition in PostgreSQL database

The programming language employed for the stated objective was HTML, CSS, JavaScript, and Bootstrap at the front end and python (Django Framework) at the back end. The back-end programming language was chosen as python because it is rich in a set of modules.

These modules made it easier to provide interfaces between the program and the Google scholar web page, the interface and the destination of the results i.e. the PostgreSQL database and the Excel sheet. Furthermore, it reduced the size of the code and helped in faster outputs with efficient results. Python also provides in-built functions for data acquisition directly from the web page. Bootstrap made a simple and user-friendly interface for input purposes. Due to these advantages which the above technologies offer, it was the go-to choice for us to make this interface. The following interface was completely built on the VSCode which can be downloaded from the code.visualstudio.com.

The platform on which the whole interface works is Windows 10 and Windows 11. The results and outcomes were stored in the PostgreSQL database and the user can download them as an Excel sheet if required. VSCode is a very simple platform for writing python code and it provides various extensions to make the tasks easier. Django is a high-level Python web framework that helps in developing secure and maintainable websites. Django takes care of much of the hassle of web development, so we just focused on writing our app. PostgreSQL comes with many features which help developers build applications, administrators protect data integrity and build fault-tolerant environments, and helps to manage data no matter how big or small the dataset. Overall, all these platforms were chosen for better connectivity.

IV. RESULTS

124 Publications Found

PUBLICATIONS

FACULTY_ID	FACULTY_NAME	TITLE	AUTHORS	PUBLICATION TYPE	NAME	PUBLISHER	PUBLISHED YEAR	ISSN	IMPACT FACTOR	CITATIONS	SCOPUS	WEB OF SCIENCE
FA1201	Dr Ravi Prakash Reddy	A Literature Survey And Comprehensive Study Of Intrusion Detection	Sravan Kumar Jonnalagadda, Ravi Prakash Reddy	-	-	-	2013	--	0.0	15	0	0
FA1201	Dr Ravi Prakash Reddy	Study And Comparison Of Non-Traditional Cloud Storage Services For High Load Text Data	L Srinivasa Rao, I Raviprakash Reddy	CONFERENCE	Proceedings Of First International Conference On Information And Communication Technology For Intelligent Systems: Volume 2	Springer International Publishing	2016	ISBN:978-3-319-30927-9	0.0	2	1	0
FA1201	Dr Ravi Prakash Reddy	Exploring The Dichotomy On Opportunities And Challenges Of Smart Technologies In Healthcare Systems	S Prabavathy, I Ravi Prakash Reddy	-	-	Academic Press	2022	--	0.0	0	1	0
FA1201	Dr Ravi Prakash Reddy	An Efficient Intrusion Detection System With Convolutional Neural Network	V Maheshwar Reddy, I Ravi Prakash Reddy, K Adi Narayana Reddy	CONFERENCE	Advances In Computational Intelligence And Informatics: Proceedings Of ICACII 2019	Springer Singapore	2020	ISBN:978-981-15-3338-9	0.0	1	1	0

Fig 5 List of publications

Update Profile | + PUBLICATION | Download

PUBLICATIONS

S. NO.	TITLE	PUBLISHED YEAR	
1	Estimation Of Human Posture Through Deep Neural Networks	2021	
2	A Complete Study On Heart Disease Detection Approach Based On Analysis Of Multiple Machine Learning Algorithms	2021	
3	Automatic Image Captioning Methodology A Tool For Visually Impaired People	2021	
4	A 3-Stage Method For Disease Detection Of Cotton Plant Leaf Using Deep Learning CNN Algorithm	2021	
5	Impact Of Communication And Coordination Factors On GSD Projects	2020	

Fig 6 Profile page

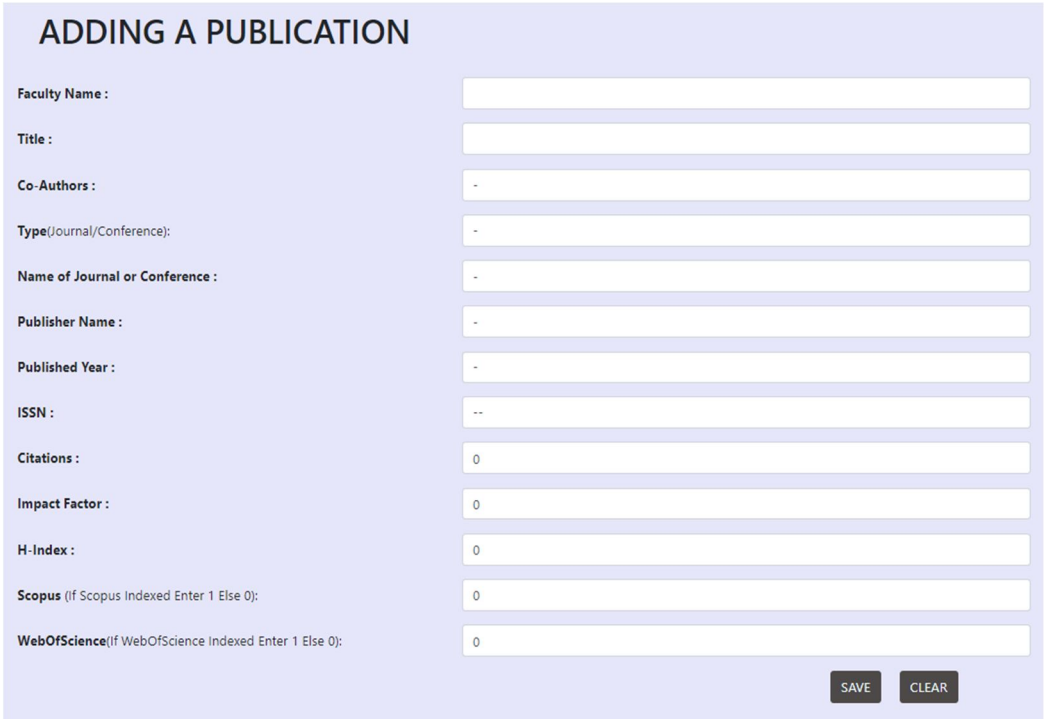


Fig 7 Adding new publication

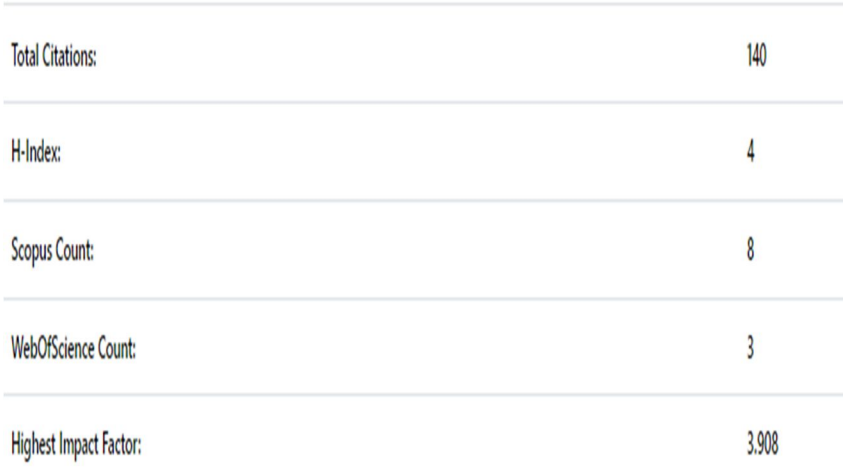


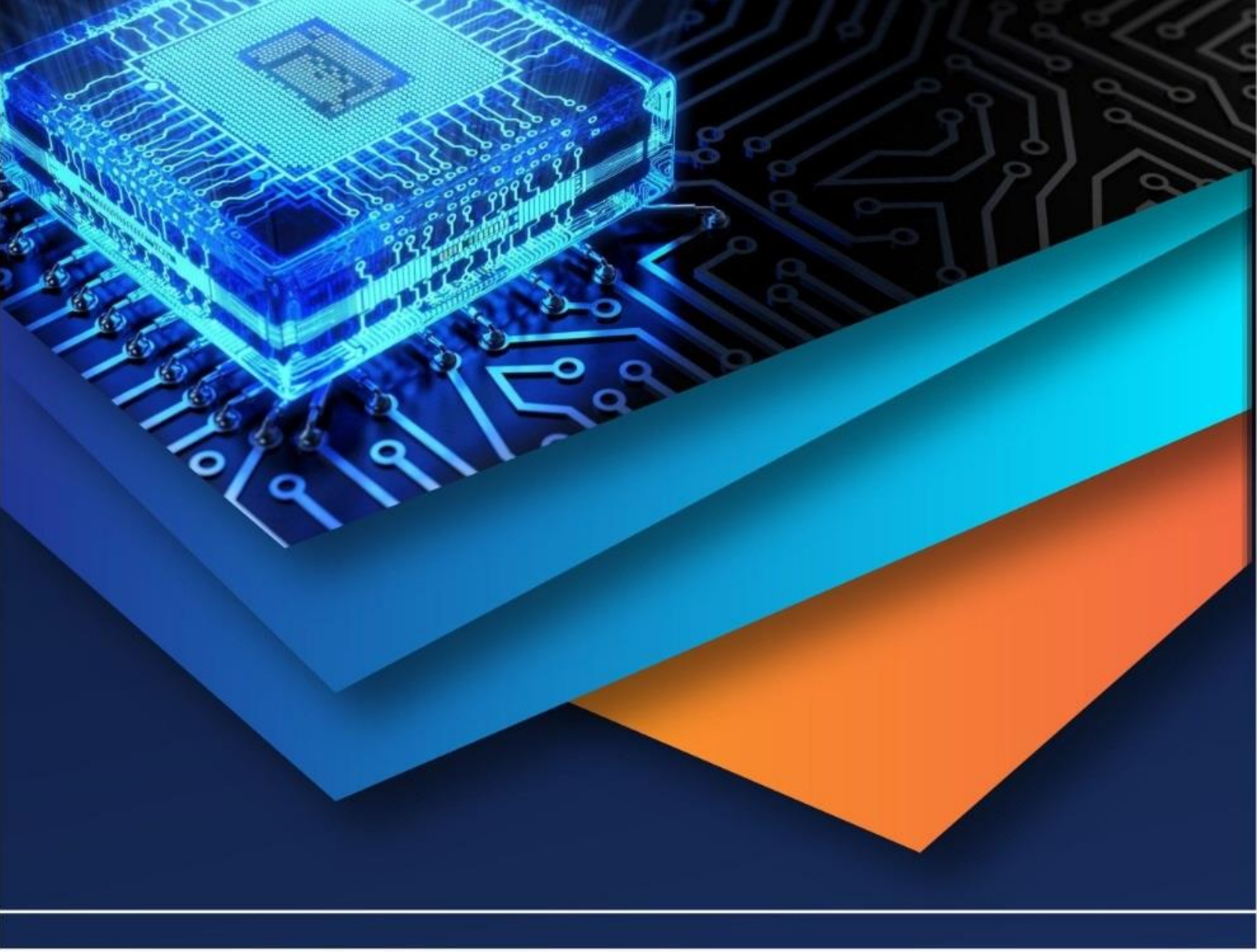
Fig 8 Summary Report

V. CONCLUSIONS & FUTURE WORK

The work as of now uses an excel sheet that stores faculty names and URLs referring to their accounts in Google scholar in order to extract the details of the publication, instead development can be made so that publications can be extracted directly using faculty names. This can be done using NLP. NLP allows creating different patterns based on their full name and university name in order to identify all the available Google Scholar accounts of a particular faculty. Then we can even extract publication details of faculty having more than one Google Scholar account. In order to check whether the publication is WebOfScience indexed the faculty ids and author-ids are collected in an excel sheet, instead development can be made so that publications are searched by opening their profiles on the WebOfScience website. An update option is provided so that the profile gets updated with new publications in Google Scholar. Manual adding and editing of publication details are provided to the faculty. As of now ISSN numbers and impact factors are extracted only for IEEE and Springer papers, instead they can be retrieved from all the papers published by various publishers.

REFERENCES

[1] D. PRATIBA, ABHAY M.S., AKHIL DUA, Giridhar K. SHANBHAG, NEEL BHANDARI, UTKARSH SINGH (2022). 'Web Scraping and Data Acquisition Using Google Scholar '. In: 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) .
[2] Đorđe Petrović, Ilja Stanišević (2017). 'Web scrapping and storing data in a database, a case study of the used cars market '. In: 2017 25th Telecommunication Forum (TELFOR).
[3] ERDINÇ UZUN (2022). 'A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages '. In: 2022 IEEE International Conference on Data Science and Information System (ICDSIS).
[4] Harsh Khatter, Dravid, Akshat Sharma, Ajay Kumar Kushwaha (2022). 'Web Scraping based Product Comparison Model for E-Commerce Websites '. In : 2022 IEEE International Conference on Data Science and Information System (ICDSIS).
[5] Ishan Adhikari Bairagi, Anchal Sharma, Bibhas Kumar Rana, Aditya Singh (2022). 'UNO: A Web Application using Django'. In: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N).
[6] I. S. Vershinin, A. R. Mustafina (2021). 'Performance Analysis of PostgreSQL, MySQL, Microsoft SQL Server Systems Based on TPC-H Tests'. In: 2021 International Russian Automation Conference (RusAutoCon).
[7] Sesha Bhargavi, V., Spandana, T. (2017). Recommendation Based P2P File Sharing on Disconnected MANET. In: Deiva Sundari, P., Dash, S., Das, S., Panigrahi, B. (eds) Proceedings of 2nd International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 467. Springer, Singapore. https://doi.org/10.1007/978-981-10-1645-5_18
[8] Velagaleti, Sesha. (2012). Design of A Scheme for Secure Routing in Mobile Ad Hoc Networks. 33-46. 10.5121/csit.2012.2104.
[9] P. L. Sebba, R. T. de Sousa, M. Holanda, A. P. F. Araújo and A. P. B. da Silva(2019). 'Database Administration: A Case Study at Public Defender of the Union in Brazil'. In: 14th Iberian Conf. on Information Systems and Technologies.
[10] [8] Vidhi Singrodia, Anirban Mitra and Subrata Paul (2019). 'A Review on Web Scrapping and its Applications'. In: 2019 International Conference on Computer Communication and Informatics (ICCCI).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY