ISSN:2302-8149
e-ISSN:2540-9719

A Web Scraper for Data Mining Purposes

[1] **Yasir Mahmood *, [2] Basim Mahmood**
[1] Computer Science Department, College of Computer Science and Mathematics University of Mosul,
41002 Mosul, Iraq
[2] ICT Research Unit, Computer Center, University of Mosul, 41002 Mosul, Iraq
[2] BioComplex Laboratory, Exeter, UK
*e-mail: yaser.ali@uomosul.edu.iq

(***received***: 15 April 2024, ***revised***: 15 May 2024, ***accepted***: 20 May 2024)

Abstract

The current revolution in technology makes data a crucial part of real-life applications due to its importance in making decisions. In the era of big data and the massive expansion of data streams on Internet networks and platforms, the process of data collection, mining, and analysis has become a not easy matter. Therefore, the presence of auxiliary applications for data mining and gathering has become a necessary need. Usually, companies offer special APIs to collect data from particular destinations, which needs a high cost. Generally, there is a severe lack in the literature in providing approaches that offer flexible, low, or free of cost tools for web scraping. Hence, this article provides a free tool that can be used for data mining and data collection purposes from the web. Specifically, an efficient Google Scholar web scraper is introduced. The extracted data can be used for analysis purposes and making decisions about an issue of interest. The proposed scraper can also be modified for crawling web links and retrieving specific data from a particular website. It can also formalize the collected data as a ready dataset to be used in the analysis phase. The efficiency of the proposed scraper is tested in terms of the time consumption, accuracy, and quality of the data collected. The findings showed that the proposed approach is highly feasible for data collection and can be adopted by data analysts.

Keywords: Web Scraping, Data Mining, Information Retrieval, Google Scholar, Complex Networks

1 Introduction

With the advent of the Internet, data has become larger and more complex. The process of analyzing and investigating data plays a significant role in making decisions. Moreover, the internet is a vast, dynamic repository of many various types of data, some beneficial and others not so helpful. It is occasionally needed to gather and retain this data for a variety of reasons, but doing that manually would probably take a very long time. Therefore, there is a critical need to automate the process of data collection to save cost and effort. Many websites offer public mechanisms (e.g., APIs) for their customers to retrieve data. Unfortunately, not all websites offer APIs for data collection, and this is where the scraper role comes in handy in the data collection process [1].

A web scraper is a program that submits an HTTP request for a certain website, waits for a legitimate response that ideally includes the HTML, and then parses the HTML for specified information [2]. Scrapers can retrieve data in a variety of forms such as text and JSON. These formats can be further reformalized to be appropriate for analysis purposes. The scraper can also be involved in formalizing the collected data as part of its tasks, which makes the process fully automated and the data analyst can directly use the output of the scraper as input of the data analysis software [3].

Many difficulties can be faced during the scraping process such as blocking the user's IP, a limited number of HTTP requests, a limited size of data collection per day, and more [4]. Therefore, developers try to use sophisticated approaches that make the scraping work smoothly. On the other hand, the scraping process can be triggered using text keywords that may exist in a particular web link or website.

2 Literature Review

The literature has a lot of works that suggest methods for web scraping, these works used a variety of approaches and techniques. The authors in [5] concentrated on a variety of web scraping-related topics, starting with a basic introduction and a quick examination of the different technologies and tools available. Hillen [6] discussed web scraping to extract huge amounts of data from various online sources. Sharma [7] looked at the concepts of web mining and its classifications. Moreover, the authors [8] presented an analysis approach to the performance of website crawlers using single and multiple indexing and algorithms.

Other studies used particular techniques for efficient scraping such as the work of Pramudita et al. [9] who used a multithreaded technique to construct the web crawler program. In [10], the authors covered the topics of web crew members, including their designs, and other difficulties that arise when search engines utilize Web crawlers.

Other recent works in the literature suggest methods that support the scraping process in many different aspects. For instance, Zhang et al. [11], proposed a dataset on documents on the web with a broad range of domains to help with training later, in addition, the authors present STAPI, for identifying part titles and prose material in HTML documents using the two-step technique.

Since Google Scholar is considered one of the main sources of bibliographic information on academic authors around the world, it is mainly used to analyze the performance of authors, institutions, and publishers, to mention but a few. The authors in [12] designed a scraper that worked to comprehend the web page's structure, and uses a designed regular expression to get certain data using API. Victoriano et al. [13] designed a scraper to retrieve bibliographic information of the faculties of Bulacan State University aiming to monitor and analyze their research activities and publications. In the same context, Rahmatulloh & Gunawan [14] designed a scraper to collect data from Google Scholar to investigate the performance of authors in terms of research publications. Their designed scraper was efficient in retrieving data successfully. In fact, there are many recent articles about scraping Google Scholar such as [15], [16], [17], [18], [19], [20]. Table 1 presents a summary of the studies considered in this section in terms of the main observed limitations.

Table 1. Literature limitations and issues.

Issue #	Study	Description
1.	[5]	Web scraping needs special programming skills as well as some practices and standardizations (a clear approach for web scraping).
2.	[6]	The scraping process may face some legal and ethical challenges.
3.	[7]	No standard tools or standardizations
4.	[8]	The main concerns were data harvest ratio and the precision of collected data.
5.	[9]	Limitations related to thread management during the collection process.
6.	[10]	Designing a web scraper depends on the purpose of the scraper and the nature of the application.
7.	[11]	The accuracy of collected data should be tested and evaluated in order to assess the efficiency of the scraper.
8.	[12]	More work is required to build an efficient scraper for data collection.
9.	[13]	There is no standard approach for scraping data.
10.	[14]	Scrapers should be executed periodically and continuously for specific targets of interest aiming to have up-to-date data.
11.	[15]	The dynamic nature of data on the web makes it difficult to get accurate insight about the data.
12.	[16]	Obstacles may be faced during the data collection process such as CAPTCHA, which should be passed during the process.
13.	[17]	Data redundancy may occur during the scraping process and data validation.
14.	[18]	Data structure and inconsistency.
15.	[19]	Source data updates during the scraping process.
16.	[20]	Natural language issues in data such as the languages that have complex scripts (e.g., Chinese, Arabic, etc.)

2.1 Literature Gap and Contribution

The deep view of the literature shows that developers used different kinds of techniques to scrape Google Scholar data and some of them are not free. This makes developers and readers confused about which method is most appropriate. In Table 1, the main observations of the literature were presented, therefore, considering these observations, this study is conducted to design a standard approach for developers to scrape Google Scholar data with no cost required. The proposed scraper is able to overcome some of the limitations of the literature. For instance, providing an efficient approach for scraping Google Scholar, no legal or ethical issues since the proposed scraper collects public data, efficient thread management strategy, and has high data collection accuracy. Hence, the contribution of this work is to propose a novel and efficient scraper for Google Scholar to efficiently retrieve information about authors, articles, institutions, journals, and publishers efficiently and with no cost required. The retrieved information can be used in any research that needs a ready-made data set in the analysis process by the researchers.

The remaining sections of this article are: Section 3 describes the design of the proposed scraper. Section 4 presents the results obtained and discussions. Finally, the whole work is concluded in Section 4.

3 Research Method

This section describes the design of the proposed scraper into two portions; first, the stages of the whole design, second, the algorithm of the proposed scraper.

3.1 Overview of the Proposed Web Scraper

The proposed Google Scholar scraper system consists of six stages shown in Figure 1. The first stage is to get the initial Uniform Resource Locator (URL) which is the seed of the scraper, and this URL is the address of the first page that we want to scrape. The second stage is the process of fetching the page that contains the information. The next stage is parsing the page and extracting the information from the parsing page. The fourth stage organizes the extracted data in a readable format. Then, the stage before the final is storing the readable data to be ready for use. The final stage is the process of getting a new URL, and this process is done either by extracting a new URL inside the fetched current page or by the next pages from the initial URL.

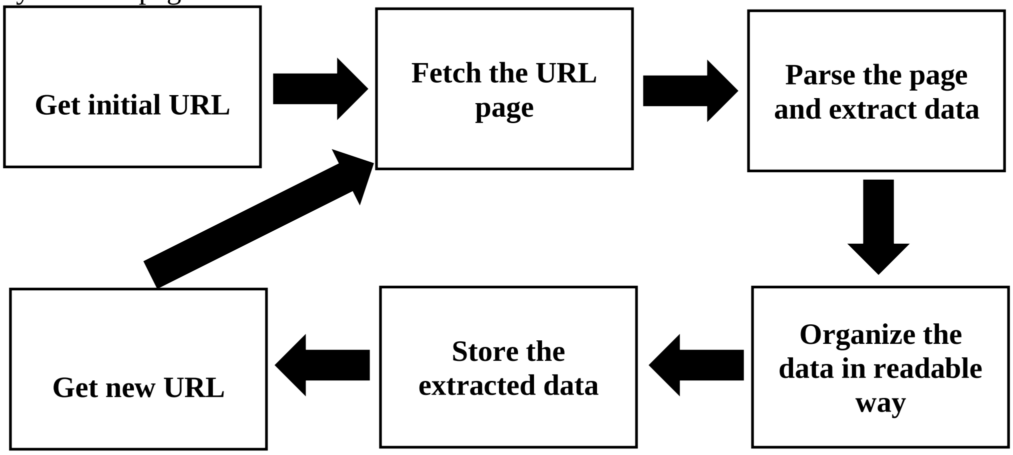


Figure 1. Stages of the proposed scraping system

3.2 The Proposed Scraping Algorithm

For easy understanding of the proposed scraper, an algorithm has been designed that simulates the work of the proposed scraper. Figure 2 shows the flowchart of the proposed system algorithm. First, we must determine the initial URL of the page that must be scrapped. This step is done by putting the initial URL to the proposed algorithm or we can generate a URL from the entry keyword.

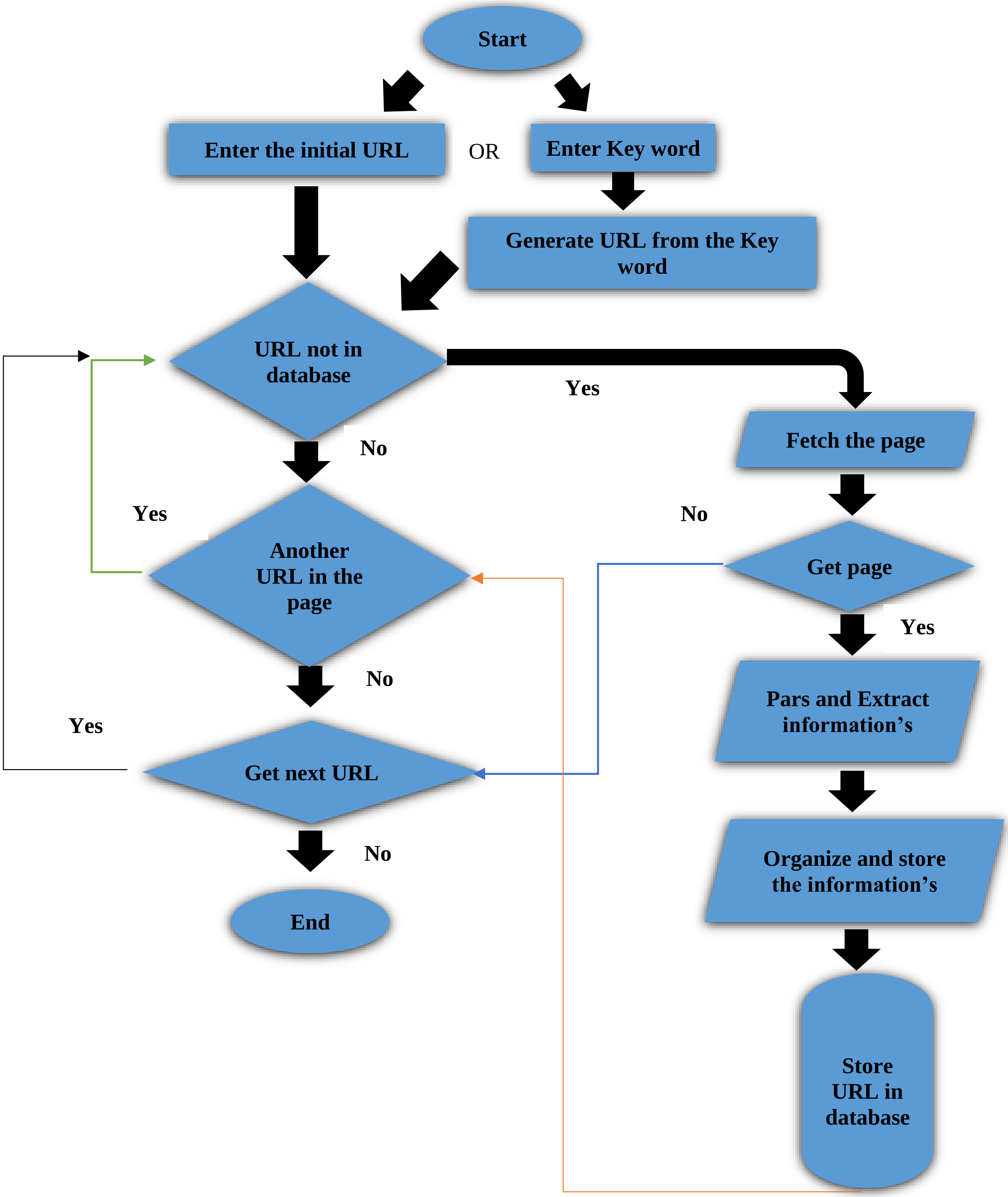


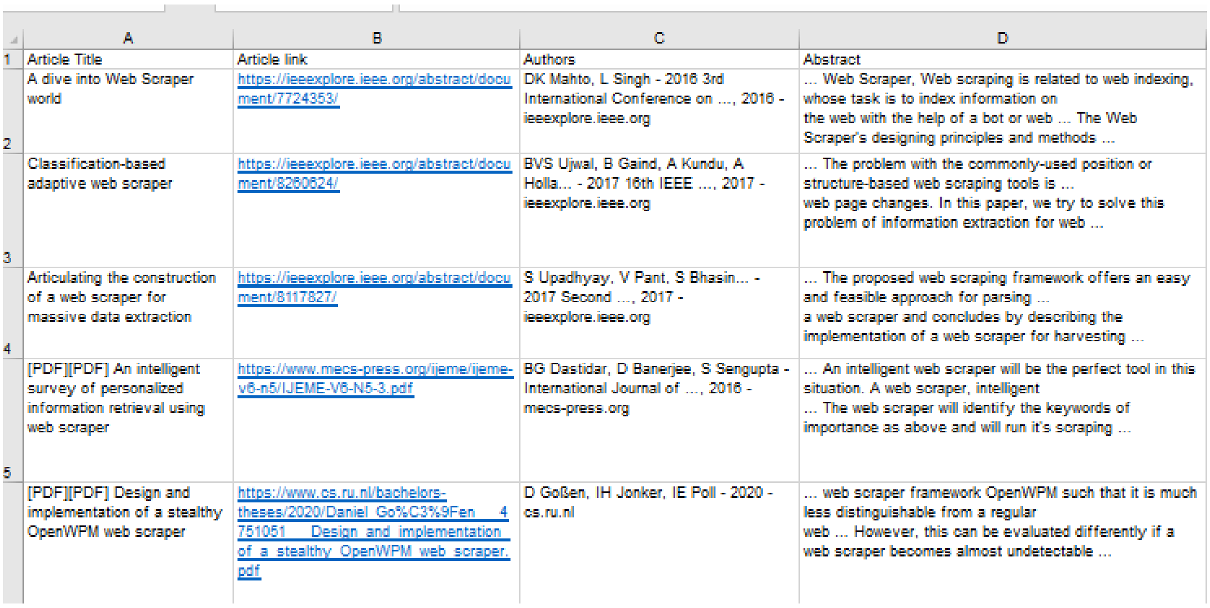
Figure 2. Flowchart for the proposed system

The next step in the proposed algorithm is to check whether the current URL address has not been visited previously to avoid entering the algorithm into an iterative loop as well as to prevent repeated scraping of pre-existing information.

ISSN:2302-8149
e-ISSN:2540-9719

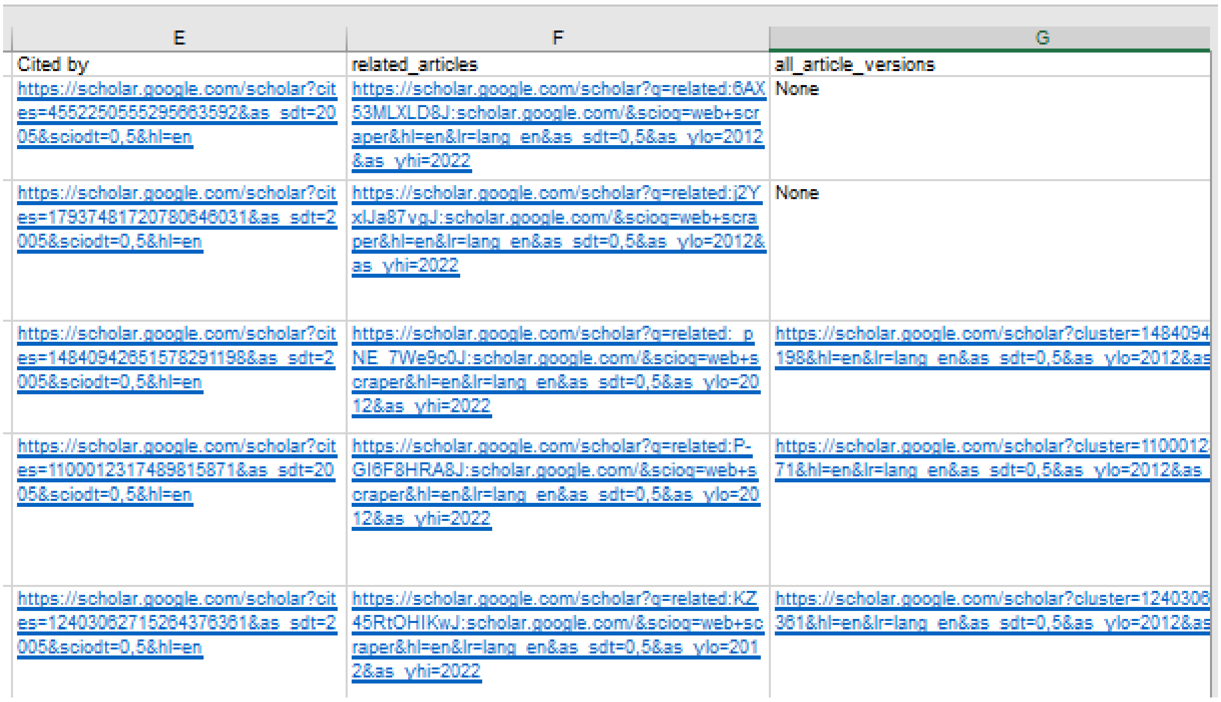
Now, if the URL is in the database, then the next step is to search for another URL on the current page, if there is another URL, then return to the previous step, or go to the next step in the proposed algorithm.

The next is, if the URL is not in the database, then the next step is to send a request to fetch the page according to the current URL, if the request gets successful and gets the page, then the next step is to parse the page to extract the information. After that, the extracted information is organized legibly, and the next step is to store the organized information for later use in research by researchers. Figure 3(A) and (B) demonstrate the sample of extracted information. The next step is to store the current URL in the database and return to the step of checking whether there is another URL in the current fetched page.



	A	B	C	D
1	Article Title	Article link	Authors	Abstract
2	A dive into Web Scraper world	https://ieeexplore.ieee.org/abstract/document/7724353/	DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org	... Web Scraper, Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web ... The Web Scraper's designing principles and methods ...
3	Classification-based adaptive web scraper	https://ieeexplore.ieee.org/abstract/document/8260624/	BVS Ujwal, B Gaind, A Kundu, A Holla... - 2017 16th IEEE ..., 2017 - ieeexplore.ieee.org	... The problem with the commonly-used position or structure-based web scraping tools is ... web page changes. In this paper, we try to solve this problem of information extraction for web ...
4	Articulating the construction of a web scraper for massive data extraction	https://ieeexplore.ieee.org/abstract/document/8117827/	S Upadhyay, V Pant, S Bhasin... - 2017 Second ..., 2017 - ieeexplore.ieee.org	... The proposed web scraping framework offers an easy and feasible approach for parsing ... a web scraper and concludes by describing the implementation of a web scraper for harvesting ...
5	[PDF][PDF] An intelligent survey of personalized information retrieval using web scraper	https://www.mecs-press.org/ijeme/ijeme-v6-n5/IJEME-V6-N5-3.pdf	BG Dastidar, D Banerjee, S Sengupta - International Journal of ..., 2016 - mecs-press.org	... An intelligent web scraper will be the perfect tool in this situation. A web scraper, intelligent ... The web scraper will identify the keywords of importance as above and will run it's scraping ...
	[PDF][PDF] Design and implementation of a stealthy OpenWPM web scraper	https://www.cs.ru.nl/bachelors-theses/2020/Daniel_Go%C3%9Fen___4751051___Design_and_implementation_of_a_stealthy_OpenWPM_web_scraper.pdf	D Goßen, IH Jonker, IE Poll - 2020 - cs.ru.nl	... web scraper framework OpenWPM such that it is much less distinguishable from a regular web ... However, this can be evaluated differently if a web scraper becomes almost undetectable ...

Figure 3. A sample of extracted information



E	F	G
Cited by	related_articles	all_article_versions
https://scholar.google.com/scholar?cites=4552250555295663592&as_sdt=2005&sciodt=0,5&hl=en	https://scholar.google.com/scholar?q=related:6AX53MLXLD8J:scholar.google.com/&scioq=web+scraper&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as_yhi=2022	None
https://scholar.google.com/scholar?cites=17937481720780646031&as_sdt=2005&sciodt=0,5&hl=en	https://scholar.google.com/scholar?q=related:j2YxlJa87vgJ:scholar.google.com/&scioq=web+scraper&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as_yhi=2022	None
https://scholar.google.com/scholar?cites=14840942651578291198&as_sdt=2005&sciodt=0,5&hl=en	https://scholar.google.com/scholar?q=related:_pNE_7We9c0J:scholar.google.com/&scioq=web+scraper&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as_yhi=2022	https://scholar.google.com/scholar?cluster=1484094198&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as
https://scholar.google.com/scholar?cites=11000123174898158071&as_sdt=2005&sciodt=0,5&hl=en	https://scholar.google.com/scholar?q=related:P-GI6F8HRA8J:scholar.google.com/&scioq=web+scraper&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as_yhi=2022	https://scholar.google.com/scholar?cluster=110001271&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as
https://scholar.google.com/scholar?cites=12403062715264376361&as_sdt=2005&sciodt=0,5&hl=en	https://scholar.google.com/scholar?q=related:KZ45RtOHIKwJ:scholar.google.com/&scioq=web+scraper&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as_yhi=2022	https://scholar.google.com/scholar?cluster=1240306361&hl=en&lr=lang_en&as_sdt=0,5&as_ylo=2012&as

Figure 3. B sample of extracted information

ISSN:2302-8149
e-ISSN:2540-9719

Now, if the request does not get successful, then the next step is to get the next URL and return to the step of checking whether the URL is in the database or not. The proposed algorithm will still work until all the pages are visited and there is no URL to get pages.

4 Results and Analysis

Designing web scrapers needs programming tools and libraries, as for the proposed scraper, some libraries were used:

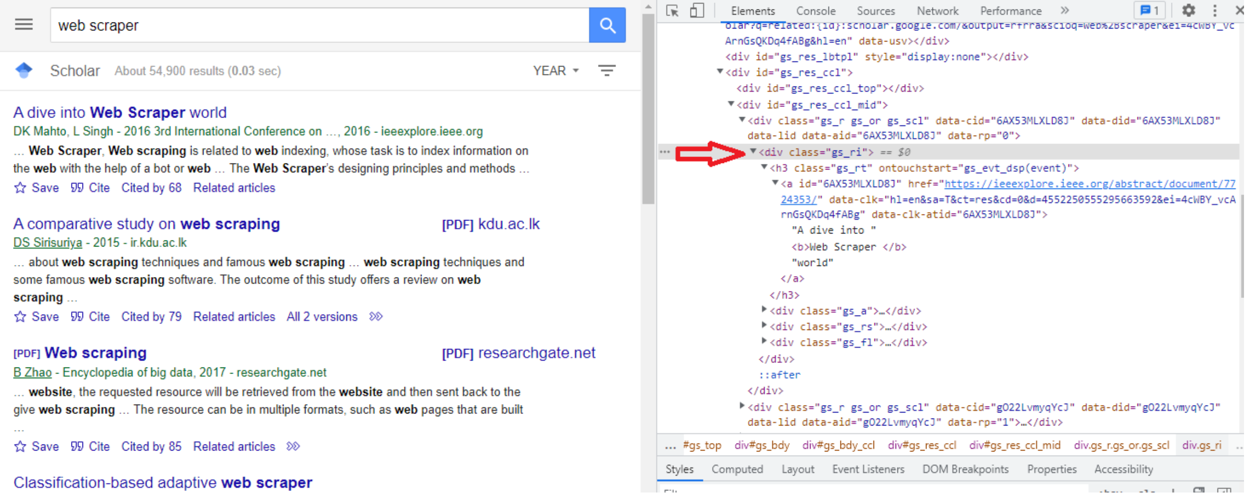
□ BeautifulSoup: is a Python language library that interprets HTML texts and retrieves data from them.

□ Lxml: is a Python language library that can be used for online scraping as well as easy management of XML and HTML files.

□ Requests: is a Python language library that is used to send requests for a web page and return the HTML form of the web page.

In general, the HTML page information is organized into divisions. Therefore, to access certain information within the HTML page, we must search for its specific division. The information below is scraped using the proposed scraper using the division tag <div>.

Article Title: the title of the paper is scraped using division '.gs_rt', as shown in Figure 4.



About 54,900 results (0.03 sec)

A dive into Web Scraper world
DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org
... Web Scraper, Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web ... The Web Scraper's designing principles and methods ...
☆ Save 99 Cite Cited by 68 Related articles

A comparative study on web scraping [PDF] kdu.ac.lk
DS Sirisuriya - 2015 - ir.kdu.ac.lk
... about web scraping techniques and famous web scraping ... web scraping techniques and some famous web scraping software. The outcome of this study offers a review on web scraping ...
☆ Save 99 Cite Cited by 79 Related articles All 2 versions

[PDF] Web scraping [PDF] researchgate.net
B Zhao - Encyclopedia of big data, 2017 - researchgate.net
... website, the requested resource will be retrieved from the website and then sent back to the give web scraping ... The resource can be in multiple formats, such as web pages that are built ...
☆ Save 99 Cite Cited by 85 Related articles

Classification-based adaptive web scraper

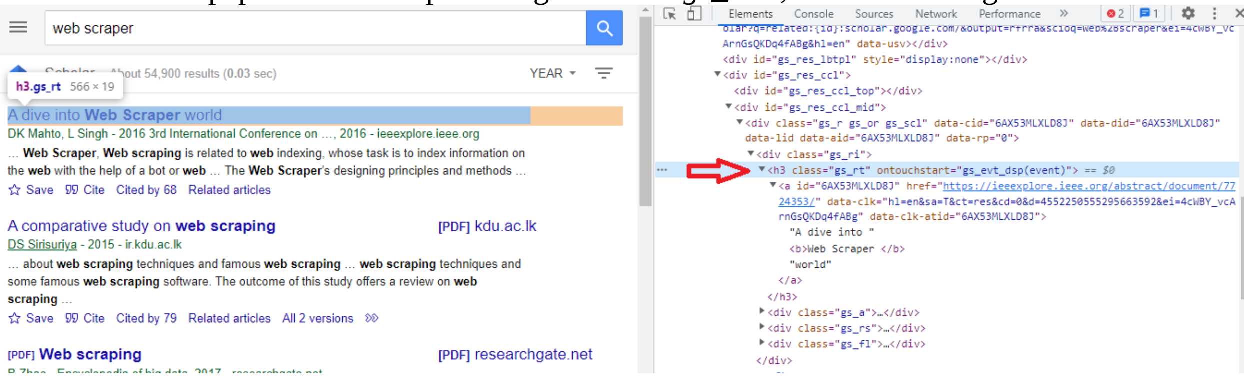
```
olar?q=related:{id}:scholar.google.com/&output=rffra&scioq=web%2Bscraper&ei=4cWBY_vC
ArnGsQKDq4fABg&hl=en" data-usv></div>
<div id="gs_res_lbtpl" style="display:none"></div>
▼<div id="gs_res_ccl">
  <div id="gs_res_ccl_top"></div>
 ▼<div id="gs_res_ccl_mid">
  ▼<div class="gs_r gs_or gs_scl" data-cid="6AX53MLXLD8J" data-did="6AX53MLXLD8J"
   data-lid data-aid="6AX53MLXLD8J" data-rp="0">
   ▼<div class="gs_ri"> == $0
    ▼<h3 class="gs_rt" ontouchstart="gs_evt_dsp(event)">
     ▼<a id="6AX53MLXLD8J" href="https://ieeexplore.ieee.org/abstract/document/77
      24353/" data-clk="hl=en&sa=T&ct=res&cd=0&d=4552250555295663592&ei=4cWBY_vcA
      rnGsQKDq4fABg" data-clk-atid="6AX53MLXLD8J">
        "A dive into "
        <b>Web Scraper </b>
        "world"
      </a>
     </h3>
    ▶<div class="gs_a">…</div>
    ▶<div class="gs_rs">…</div>
    ▶<div class="gs_fl">…</div>
    </div>
    ::after
   </div>
  ▶<div class="gs_r gs_or gs_scl" data-cid="gO22LvmyqYcJ" data-did="gO22LvmyqYcJ"
   data-lid data-aid="gO22LvmyqYcJ" data-rp="1">…</div>
```

... #gs_top div#gs_bdy div#gs_bdy_ccl div#gs_res_ccl div#gs_res_ccl_mid div.gs_r.gs_or.gs_scl div.gs_ri ...

Figure 4. Extraction of the article title

The other piece of information that was retrieved:

Article link: the paper link is scraped using division *'.gs_rt a'*, as shown in Figure 5.



About 54,900 results (0.03 sec)

A dive into Web Scraper world
DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org
... Web Scraper, Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web ... The Web Scraper's designing principles and methods ...
☆ Save 99 Cite Cited by 68 Related articles

A comparative study on web scraping [PDF] kdu.ac.lk
DS Sirisuriya - 2015 - ir.kdu.ac.lk
... about web scraping techniques and famous web scraping ... web scraping techniques and some famous web scraping software. The outcome of this study offers a review on web scraping ...
☆ Save 99 Cite Cited by 79 Related articles All 2 versions

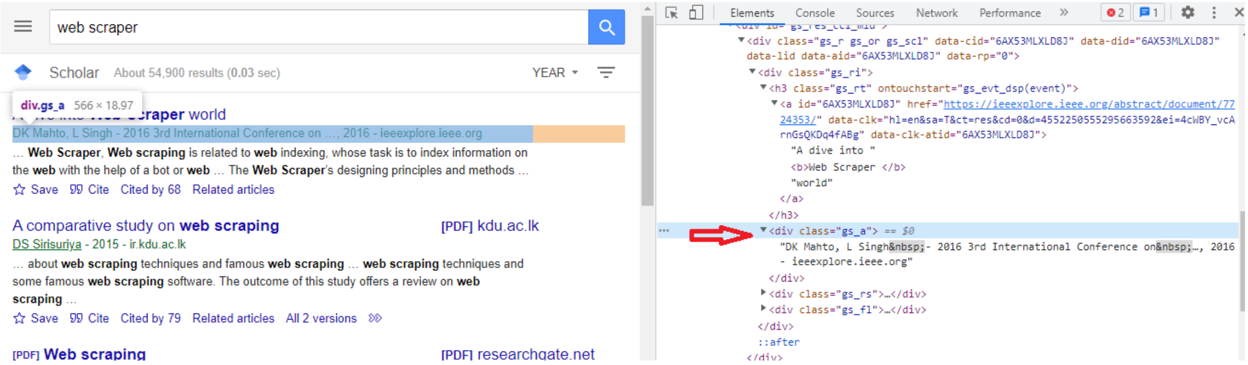
[PDF] Web scraping [PDF] researchgate.net

```
olar?q=related:{id}:scholar.google.com/&output=rffra&scioq=web%2Bscraper&ei=4cWBY_vC
ArnGsQKDq4fABg&hl=en" data-usv></div>
<div id="gs_res_lbtpl" style="display:none"></div>
▼<div id="gs_res_ccl">
  <div id="gs_res_ccl_top"></div>
 ▼<div id="gs_res_ccl_mid">
  ▼<div class="gs_r gs_or gs_scl" data-cid="6AX53MLXLD8J" data-did="6AX53MLXLD8J"
   data-lid data-aid="6AX53MLXLD8J" data-rp="0">
   ▼<div class="gs_ri">
    ▼<h3 class="gs_rt" ontouchstart="gs_evt_dsp(event)"> == $0
     ▼<a id="6AX53MLXLD8J" href="https://ieeexplore.ieee.org/abstract/document/77
      24353/" data-clk="hl=en&sa=T&ct=res&cd=0&d=4552250555295663592&ei=4cWBY_vcA
      rnGsQKDq4fABg" data-clk-atid="6AX53MLXLD8J">
        "A dive into "
        <b>Web Scraper </b>
        "world"
      </a>
     </h3>
    ▶<div class="gs_a">…</div>
    ▶<div class="gs_rs">…</div>
    ▶<div class="gs_fl">…</div>
    </div>
```

Figure 5. Extraction of article link

Also, the authors' information was retrieved as follows:

Authors: the information of the authors of the article can be scraped using division *'.gs_a'*, as shown in Figure 6.



web scraper

Scholar About 54,900 results (0.03 sec) YEAR

div.gs_a 566 × 18.97 raper world

DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org

... **Web Scraper**, **Web scraping** is related to **web** indexing, whose task is to index information on the **web** with the help of a bot or **web** ... The **Web Scraper**'s designing principles and methods ...

☆ Save 99 Cite Cited by 68 Related articles

A comparative study on **web scraping** [PDF] kdu.ac.lk

DS Sirisuriya - 2015 - ir.kdu.ac.lk

... about **web scraping** techniques and famous **web scraping** ... **web scraping** techniques and some famous **web scraping** software. The outcome of this study offers a review on **web scraping** ...

☆ Save 99 Cite Cited by 79 Related articles All 2 versions

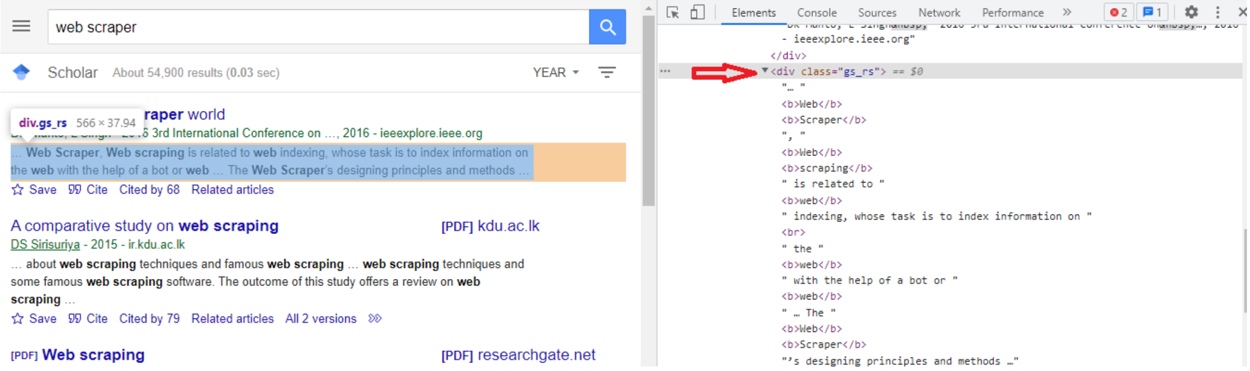
[PDF] **Web scraping** [PDF] researchgate.net

```
▼<div class="gs_r gs_or gs_scl" data-cid="6AX53MLXLD8J" data-did="6AX53MLXLD8J"
 data-lid data-aid="6AX53MLXLD8J" data-rp="0">
  ▼<div class="gs_ri">
    ▼<h3 class="gs_rt" ontouchstart="gs_evt_dsp(event)">
      ▼<a id="6AX53MLXLD8J" href="https://ieeexplore.ieee.org/abstract/document/77
       24353/" data-clk="hl=en&sa=T&ct=res&cd=0&d=4552250555295663592&ei=4cWBY_vcA
       rnGsQKDq4fABg" data-clk-atid="6AX53MLXLD8J">
         "A dive into "
         <b>Web Scraper </b>
         "world"
       </a>
     </h3>
    ▼<div class="gs_a"> == $0
       "DK Mahto, L Singh&nbsp;- 2016 3rd International Conference on&nbsp;…, 2016
       - ieeexplore.ieee.org"
     </div>
    ▶<div class="gs_rs">…</div>
    ▶<div class="gs_fl">…</div>
   </div>
   ::after
 </div>
```

Figure 6. Extraction of authors' information

Other information such as abstract, cited by, related articles, and article versions were also scraped as follows:

Abstract: the abstract of the article can be scraped using the division *'.gs_rs'*, as shown in Figure 7.



web scraper

Scholar About 54,900 results (0.03 sec) YEAR

div.gs_rs 566 × 37.94 raper world

3rd International Conference on ..., 2016 - ieeexplore.ieee.org

... **Web Scraper**, **Web scraping** is related to **web** indexing, whose task is to index information on the **web** with the help of a bot or **web** ... The **Web Scraper**'s designing principles and methods ...

☆ Save 99 Cite Cited by 68 Related articles

A comparative study on **web scraping** [PDF] kdu.ac.lk

DS Sirisuriya - 2015 - ir.kdu.ac.lk

... about **web scraping** techniques and famous **web scraping** ... **web scraping** techniques and some famous **web scraping** software. The outcome of this study offers a review on **web scraping** ...

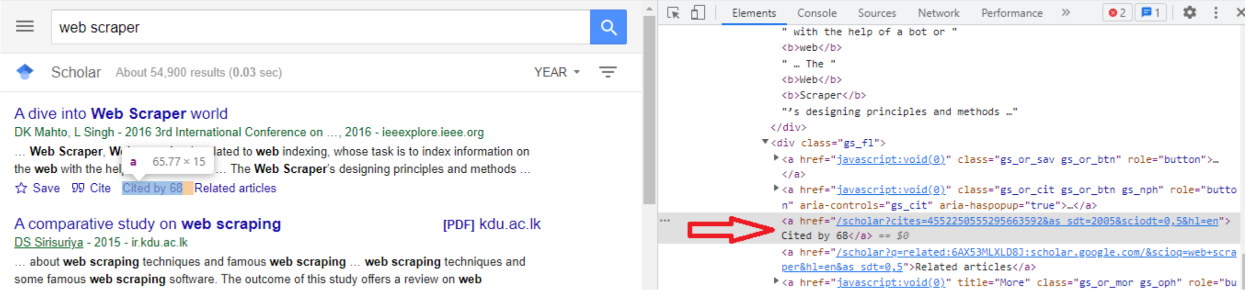
☆ Save 99 Cite Cited by 79 Related articles All 2 versions

[PDF] **Web scraping** [PDF] researchgate.net

```
       - ieeexplore.ieee.org"
     </div>
    ▼<div class="gs_rs"> == $0
       "… "
       <b>Web</b>
       <b>Scraper</b>
       ", "
       <b>Web</b>
       <b>scraping</b>
       " is related to "
       <b>web</b>
       " indexing, whose task is to index information on "
       <br>
       " the "
       <b>web</b>
       " with the help of a bot or "
       <b>web</b>
       " … The "
       <b>Web</b>
       <b>Scraper</b>
       "'s designing principles and methods …"
```

Figure 7. Extracting of the abstract of an article

Cited by: also, we can scrape a link representing the information about who cited this article by using the division *'#gs_res_ccl_mid .gs_nph+ a'*, as shown in Figure 8.



web scraper

Scholar About 54,900 results (0.03 sec) YEAR

A dive into **Web Scraper** world

DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org

... **Web Scraper**, **W** a 65.77 × 15 lated to **web** indexing, whose task is to index information on the **web** with the hel ... The **Web Scraper**'s designing principles and methods ...

☆ Save 99 Cite Cited by 68 Related articles

A comparative study on **web scraping** [PDF] kdu.ac.lk

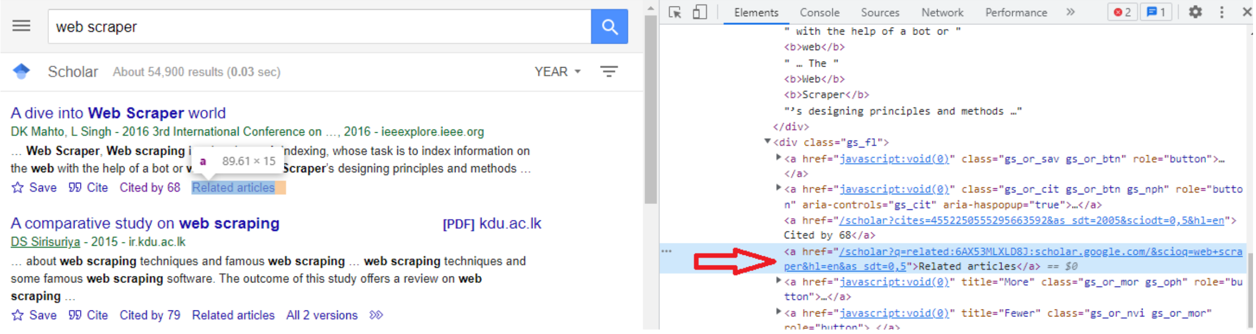
DS Sirisuriya - 2015 - ir.kdu.ac.lk

... about **web scraping** techniques and famous **web scraping** ... **web scraping** techniques and some famous **web scraping** software. The outcome of this study offers a review on **web**

```
       " with the help of a bot or "
       <b>web</b>
       " … The "
       <b>Web</b>
       <b>Scraper</b>
       "'s designing principles and methods …"
     </div>
    ▼<div class="gs_fl">
      ▶<a href="javascript:void(0)" class="gs_or_sav gs_or_btn" role="button">…
       </a>
      ▶<a href="javascript:void(0)" class="gs_or_cit gs_or_btn gs_nph" role="butto
       n" aria-controls="gs_cit" aria-haspopup="true">…</a>
       <a href="/scholar?cites=4552250555295663592&as_sdt=2005&sciodt=0,5&hl=en">
       Cited by 68</a> == $0
       <a href="/scholar?q=related:6AX53MLXLD8J:scholar.google.com/&scioq=web+scra
       per&hl=en&as_sdt=0,5">Related articles</a>
      ▶<a href="javascript:void(0)" title="More" class="gs_or_mor gs_oph" role="bu
```

Figure 8. Extracting of cited by

related_articles: the related articles link can be scraped using the division *'a:nth-child(4)'*, as shown in Figure 9.



web scraper

Scholar About 54,900 results (0.03 sec) YEAR

A dive into **Web Scraper** world

DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org

... **Web Scraper**, **Web scraping** i a 89.61 × 15 ndexing, whose task is to index information on the **web** with the help of a bot or **v** **Scraper**'s designing principles and methods ...

☆ Save 99 Cite Cited by 68 Related articles

A comparative study on **web scraping** [PDF] kdu.ac.lk

DS Sirisuriya - 2015 - ir.kdu.ac.lk

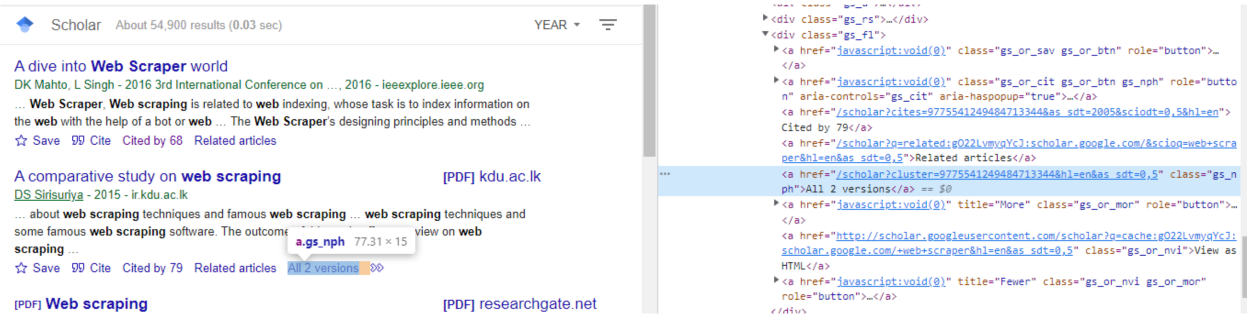
... about **web scraping** techniques and famous **web scraping** ... **web scraping** techniques and some famous **web scraping** software. The outcome of this study offers a review on **web scraping** ...

☆ Save 99 Cite Cited by 79 Related articles All 2 versions

```
       " with the help of a bot or "
       <b>web</b>
       " … The "
       <b>Web</b>
       <b>Scraper</b>
       "'s designing principles and methods …"
     </div>
    ▼<div class="gs_fl">
      ▶<a href="javascript:void(0)" class="gs_or_sav gs_or_btn" role="button">…
       </a>
      ▶<a href="javascript:void(0)" class="gs_or_cit gs_or_btn gs_nph" role="butto
       n" aria-controls="gs_cit" aria-haspopup="true">…</a>
       <a href="/scholar?cites=4552250555295663592&as_sdt=2005&sciodt=0,5&hl=en">
       Cited by 68</a>
       <a href="/scholar?q=related:6AX53MLXLD8J:scholar.google.com/&scioq=web+scra
       per&hl=en&as_sdt=0,5">Related articles</a> == $0
      ▶<a href="javascript:void(0)" title="More" class="gs_or_mor gs_oph" role="bu
       tton">…</a>
      ▶<a href="javascript:void(0)" title="Fewer" class="gs_or_nvi gs_or_mor"
       role="button">…</a>
```

Figure 9. Extracting of related articles

all_article_versions: finally, we can scrape the versions of the article by using the *division* *'a~ a+ .gs_nph'*, as shown in Figure 10.



Scholar About 54,900 results (0.03 sec) YEAR

A dive into **Web Scraper** world
DK Mahto, L Singh - 2016 3rd International Conference on ..., 2016 - ieeexplore.ieee.org
... **Web Scraper**, **Web scraping** is related to **web** indexing, whose task is to index information on the **web** with the help of a bot or **web** ... The **Web Scraper**'s designing principles and methods ...
☆ Save 99 Cite Cited by 68 Related articles

A comparative study on **web scraping** [PDF] kdu.ac.lk
DS Sirisuriya - 2015 - ir.kdu.ac.lk
... about **web scraping** techniques and famous **web scraping** ... **web scraping** techniques and some famous **web scraping** software. The outcome view on **web** **scraping** ...
☆ Save 99 Cite Cited by 79 Related articles All 2 versions

a.gs_nph 77.31 × 15

[PDF] **Web scraping** [PDF] researchgate.net

```
 ▶<div class="gs_rs">…</div>
 ▼<div class="gs_fl">
   ▶<a href="javascript:void(0)" class="gs_or_sav gs_or_btn" role="button">…
    </a>
   ▶<a href="javascript:void(0)" class="gs_or_cit gs_or_btn gs_nph" role="butto
    n" aria-controls="gs_cit" aria-haspopup="true">…</a>
    <a href="/scholar?cites=9775541249484713344&as_sdt=2005&sciodt=0,5&hl=en">
    Cited by 79</a>
    <a href="/scholar?q=related:gO22LvmyqYcJ:scholar.google.com/&scioq=web+scra
    per&hl=en&as_sdt=0,5">Related articles</a>
...   <a href="/scholar?cluster=9775541249484713344&hl=en&as_sdt=0,5" class="gs_n
    ph">All 2 versions</a> == $0
   ▶<a href="javascript:void(0)" title="More" class="gs_or_mor" role="button">…
    </a>
    <a href="http://scholar.googleusercontent.com/scholar?q=cache:gO22LvmyqYcJ:
    scholar.google.com/+web+scraper&hl=en&as_sdt=0,5" class="gs_or_nvi">View as
    HTML</a>
   ▶<a href="javascript:void(0)" title="Fewer" class="gs_or_nvi gs_or_mor"
    role="button">…</a>
 </div>
```

Figure 10. Extracting of the versions of an article

Using the proposed algorithm, developers can get huge information on specific topics determined by the URL or by keyword, this information is useful for other researchers to use in their research and analysis. Furthermore, this paper considers Google Scholar as the main source of data and uses the proposed web scraper with input keywords to get data of interest. The scraper in this paper goes into the depth of the pages, by finding URL addresses to information not available on the current page. The first depth of scraping is collecting direct information that is available on the page, and the second depth of scraping is the collection of the URL addresses from the current page and using them to collect unavailable information on the current page.

The proposed scraper is able to collect huge information from Google Scholar and arrange this information in a way that is easy to manage. Also, the proposed scraper can support researchers in collecting a dataset for any topic in any research area of interest by setting the appropriate keyword, as it consumes the time of the data collection process. Finally, the proposed scraper is considered economically sufficient since it does not require any cost for collecting the scraped links. The proposed crawler can fetch all information for approximately 100 articles every 2 seconds. A delay of two seconds was made to avoid the ban from fetching information from Google Scholar, as well as to solve the problem of Captcha. Moreover, when it is required to collect significant big data, some techniques are available in the literature that can speed up the collection process especially if it is performed completely in the cloud [21], [22], [23].

Finally, Table 2 presents the literature limitations that have been solved/unsolved using the proposed approach. However, the issues that have not yet been solved, are considered ongoing and left as a future work of this study.

Table 2. Issues solved/unsolved by the proposed scraper

Issue #	Solved (S)/ Unsolved (U)	Method
1.	S	Design a step-by-step clear approach.
2.	S	Collecting public data.
3.	S	Using well-defined tools (libraries and scripts)
4.	S	The collected data was manually tested.
5.	S	Coding well-designing thread approach.
6.	S	The target was Google Scholar data and the proposed scraper was sufficient.
7.	S	Sample data was manually counted and compared to the collected.
8.	S	The proposed scraper includes tools and techniques selected after many tries.
9.	S	For Google Scholar, the proposed approach can be considered a standard.
10.	U	The proposed scrapers should be executed periodically and continuously to get up-to-date data.
11.	U	Google Scholar data is dynamic.
12.	S	Addressed

13.	Partially	This issue is considered a limitation of the proposed scraper when having double profiles for the same author in Google Scholar.
14.	S	Google Scholar data is well-structured.
15.	U	At the time the scraper gets a profile, any update will not be considered, which is also a limitation of the proposed scraper.
16.	U	A limitation of the proposed scraper.

5 Conclusion

This work suggested a web scraper for web scraping and data mining. The proposed scraper could crawl web links and retrieve specific data from a particular website. It could also formalize the collected data as a ready dataset to be used in the analysis phase. Moreover, the case study considered in this work was Google Scholar, therefore, the proposed scraper, was mainly used to collect bibliography information from Google Scholar and arrange this information in a way that is easy to deal with and managed by researchers. The proposed scraper is adjustable to be sufficient in collecting data about a particular author, institution, publisher, or journal, this can be performed easily by setting the keywords of interest. Additionally, the proposed scraper worked in two depths, which means it can retrieve the information included in the links available on the scraped web page. The design approach of the proposed scraper was carefully performed aiming to have an efficient data collection tool in terms of time consumption, data accuracy, and the quality of data. The proposed scraper was able to overcome many of the limitations in the literature. However, some limitations need to be given more attention in the next version of the proposed scraper.

As future work, it is planned to overcome the limitations mentioned in Table 2. It is also planned to design a frontend (e.g., special purpose website) that includes the proposed scraper and makes it available for public use. The other aspect that needs more improvement is enabling the proposed scraper to work in 3 or 4 depths as required.

Acknowledgement

The authors would appreciate the support provided by the Computer Science Department/ College of Computer Science and Mathematics/ University of Mosul during the steps of this work**.**

Reference

[1] J. Lin, “A proposed conceptual framework for a representational approach to information retrieval,” *ACM SIGIR Forum*, vol. 55, no. 2, 2021, doi: 10.1145/3527546.3527552.

[2] M. A. Khder, “Web scraping or web crawling: State of art, techniques, approaches and application,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, 2021, doi: 10.15849/ijasca.211128.11.

[3] V. Morina and S. Sejdiu, “Evaluating and comparing web scraping tools and techniques for data collection.,” *UBT Knowl. Cent.*, 2022.

[4] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso, and S. N. Mbaye, “Web Scraping: State-of-the-Art and Areas of Application,” in *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 2019. doi: 10.1109/BigData47090.2019.9005594.

[5] V. Singrodia, A. Mitra, and S. Paul, “A Review on Web Scrapping and its Applications,” in *2019 International Conference on Computer Communication and Informatics, ICCCI 2019*, 2019. doi: 10.1109/ICCCI.2019.8821809.

[6] J. Hillen, “Web scraping for food price research,” *Br. Food J.*, vol. 121, no. 12, 2019, doi: 10.1108/BFJ-02-2019-0081.

[7] S. S and K. N, “WEB SCRAPPING TOOLS,” *J. Anal. Comput.*, 2019.

[8] A. K. Sharma, V. Shrivastava, and H. Singh, “Experimental performance analysis of web crawlers using single and Multi-Threaded web crawling and indexing algorithm for the application of smart web contents,” in *Materials Today: Proceedings*, 2020. doi: 10.1016/j.matpr.2020.06.596.

[9] Y. D. Pramudita, D. R. Anamisa, S. S. Putro, and M. A. Rahmawanto, “Extraction System Web Content Sports New Based on Web Crawler Multi Thread,” in *Journal of Physics: Conference Series*, 2020. doi: 10.1088/1742-6596/1569/2/022077.

[10] P. Ghantasala, V. Chinnam, R. Karukuri, and S. Sripad, "Range of web Crawling from HTTP Parse and HTML Requests as Static Digreph and Web Pages," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3669657.

[11] N. Zhang, S. Wilson, and P. Mitra, "STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents," in *2022 Language Resources and Evaluation Conference, LREC 2022*, 2022.

[12] Q. Niu *et al.*, "Web Scraping Tool For Newspapers And Images Data Using Jsonify," *J. Appl. Sci. Eng.*, vol. 26, no. 4, 2023, doi: 10.6180/jase.202304_26(4).0002.

[13] R. L. A. Victoriano, J. M., Pulumbarit, J. P., Lacatan, L. L., Salivio, R. A. S., & Barawid, "Data Analysis of Bulacan State University Faculty Scientific Publication Based on Google Scholar using Web Data Scraping Technique.," *Int. J. Comput. Sci. Res.*, 2022.

[14] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar," *Indones. J. Inf. Syst.*, vol. 2, no. 2, 2020, doi: 10.24002/ijis.v2i2.3029.

[15] J. Buyuklieva, Boyana and Raimbault, "Estimating Bibliometric Links using Google Scholar: A Semi-Systematic Literature Mapping of Migration and Housing.," *arXiv Prepr. arXiv2302.10791*, 2023.

[16] D. Murillo, D. Saavedra, and R. Zapata, "Web application in Shiny for the extraction of data from profiles in Google Scholar," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2022. doi: 10.18687/LACCEI2022.1.1.235.

[17] N. Sabah, M. M. Khan, R. Talib, M. Anwar, M. S. A. Malik, and P. Nohuddin, "Google Scholar University Ranking Algorithm to Evaluate the Quality of Institutional Research," *Comput. Mater. \& Contin.*, vol. 75, no. 3, pp. 4955--4972, 2023.

[18] N. A. Sultan and D. B. Abdullah, "Scraping Google Scholar Data Using Cloud Computing Techniques," in *2022 8th International Conference on Contemporary Information Technology and Mathematics, ICCITM 2022*, 2022. doi: 10.1109/ICCITM56309.2022.10032044.

[19] B. Soriano, C. Isaac, A. Bautista, J. López-Ramírez, and Misael, "Obtención de una base de datos de perfiles de investigadores en Google Scholar basado en web scraping," *JÓVENES EN LA Cienc.*, vol. 18, pp. 1–3, 2022.

[20] M. R. Rafsanjani, "ScrapPaper: A web scrapping method to extract journal information from PubMed and Google Scholar search result using Python.," *bioRxiv*, 2022.

[21] B. Mahmood and Y. Mahmood, "Network-Based Method for Dynamic Burden-Sharing in the Internet of Things (IoT)," in *Communications in Computer and Information Science*, 2022. doi: 10.1007/978-3-030-97255-4_6.

[22] N. A. Sultan, B. Mahmood, K. H. Thanoon, and D. S. Khadhim, "Network Centralities-Based Approach for Evaluating Interdisciplinary Collaboration," in *Proceedings of the 6th International Engineering Conference "'Sustainable Technology and Development'", IEC 2020*, 2020. doi: 10.1109/IEC49899.2020.9122915.

[23] K. H. Thanoon, N. A. Sultan, B. Mahmood, and D. S. Kadhim, "Measuring scientific collaboration in co-authorship networks," *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, 2021, doi: 10.11591/IJAI.V10.I4.PP1103-1114.