Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal.768-773
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Automated Google Scholar Crawling with a Web-Based Tool for Publication Data Management

Yudha Islami Sulistya[a], Ariq Cahya Wardhana[b], Maie Istighosah[c] Arif Riyandi[d]
[a] Rekayasa Perangkat Lunak, Telkom University, Purwokerto, Indonesia, yudhaislami@telkomuniversity.ac.id
[b] Rekayasa Perangkat Lunak, Telkom University, Purwokerto, Indonesia, ariqcahya@telkomuniversity.ac.id
[c] Teknik Informatika, Telkom University, Purwokerto, Indonesia, maieistigh@telkomuniversity.ac.id
[d] Sistem Informasi, Telkom University, Purwokerto, Indonesia, arifriyandi@telkomuniversity.ac.id

Submitted: 14-08-2024, Reviewed: 26-08-2024, Accepted 03-10-2024
https://doi.org/10.47233/jteksis.v6i4.1604

Abstract
The rapid growth of academic publications presents challenges in efficiently managing and utilizing publication data, particularly from platforms like Google Scholar. This research addresses these challenges by developing an automated web-based tool designed to extract and manage publication data. The primary objectives are to create a tool that simplifies data crawling from Google Scholar and manages the resulting data efficiently. The methodology involves developing a crawling algorithm with simple and detailed modes, integrating it into a user-friendly web interface, and conducting performance tests on datasets of varying sizes. Analytical techniques used include testing execution times for different dataset sizes and modes. The results demonstrate that the tool efficiently handles both small and large datasets, with simple mode completing tasks faster but with less detail, while detailed mode provides comprehensive data but requires longer processing times. This tool significantly enhances the efficiency of data collection processes, offering researchers flexibility in balancing speed and depth of information according to their needs. The research contributes a practical solution for managing scientific publication data, aiding researchers in effectively accessing, organizing, and analyzing information from Google Scholar.
Keywords: *google scholar, automated web crawling, data management, web-based tool*

This work is licensed under Creative Commons Attribution License 4.0 CC-BY International license

INTRODUCTION

In today's digital era, the growth of scientific publications is occurring rapidly and exponentially. This is due to the increasing number of researchers and academic institutions contributing to various disciplines. As a result, managing and utilizing publication data has become increasingly complex. Data plays a crucial role in research [1], serving as a valuable asset that can be leveraged to drive evidence-based decision-making, enhance the accuracy and efficiency of research processes, and provide deeper insight into the phenomena being studied. Google Scholar, as one of the most widely used academic search engines, provides access to millions of scholarly articles, theses, books, and patents from various sources. However, the ability to efficiently access, collect, and manage data from Google Scholar remains a challenge for many researchers, especially on a large scale [2]–[4]

Automated web crawling has become one of the solutions to address this issue, enabling fast and accurate data extraction from various online sources [5]. The use of web-based tools for crawling data from Google Scholar allows researchers to automate the data collection process, which not only saves time but also minimizes human error [6]. furthermore, effective management of publication data is crucial to support further data analysis and data-driven decision-making [7].

The urgency of this research arises from the need to develop a tool capable of handling the vast volume of publication data available on Google Scholar. Previous studies have highlighted the importance of developing tools that can integrate automated data extraction processes with user-friendly interfaces [8], [9]. his tool is expected to assist researchers in efficiently managing publication data and facilitate access to relevant information for further research.

In an effort to address this problem, this research aims to develop a web-based tool that automatically crawls data from Google Scholar and manages the resulting publication data. The methods used in this study include the development of an efficient crawling algorithm, integration with a web-based user interface, and performance testing of the tool on various dataset sizes [10]. It is expected that this tool will not only enhance the efficiency of data collection processes but also expand researchers' capacity to manage and analyze scientific publication data [11].

Previous research has shown that automation in managing publication data can significantly reduce manual workload and increase accuracy in data

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal.768-773
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

collection [12]–[14]. Additionally, the use of web-based technology allows for easier and broader access for users, which in turn can enhance collaboration and information distribution within the scientific community [15], [16].

Therefore, this research is important as it provides a tangible contribution to the development of technology that supports the management of scientific publication data. Moreover, the results of this research are expected to be widely used by researchers and academic institutions to improve the effectiveness of publication management and accelerate the research process.

RESEARCH METHOD

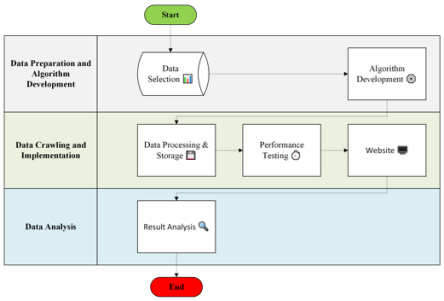

Figure 1. Research Methodology

As illustrated in Figure 1, this research is divided into three main phases: Data Preparation and Algorithm Development, Implementation and Data Processing, and Data Analysis and Reporting. In the first phase, the dataset selection is carried out by identifying and determining the URLs from Google Scholar to be crawled, followed by the development of an effective algorithm to execute the crawling process in both simple and detailed modes. The second phase focuses on processing the crawled data, where the data is converted into CSV format and, if necessary, organized into a ZIP file for download through the website. Additionally, performance testing of the tool is conducted on various dataset sizes to ensure its efficiency. The final phase involves analyzing the results obtained from the performance testing. This structured approach allows the research to proceed systematically, ensuring that all critical aspects of the development and implementation of this crawling tool function effectively.

2.1. Data Selection

In this study, data was collected from two author profiles selected from Google Scholar. Each profile was chosen based on the relevance of their publications to the related research field. The first profile contains 5 published articles, while the second profile includes 41 published articles. The selection of two datasets with different sizes aims to test the performance of the crawling tool on both small and large amounts of data, thereby allowing for a comparison of speed and accuracy in the data extraction process

2.2. Algorithm Development

In the development of the algorithm for this research, a pagination-based crawling approach was employed, which is commonly used for data extraction from websites that present data in a paginated format. The algorithm not only focuses on iteratively retrieving data from each page but also implements both simple and detailed crawling modes, allowing users to select the depth of data extraction, as shown in Table 1. References to the use of similar crawling techniques can be found in studies such as the one conducted [17], where they developed a crawling algorithm to collect data from academic websites with a similar pagination structure.

Table 1. Data Crawling Algorithm Code

Algorithm: Data Crawling Algorithm

function: crawl_pages(url, start, crawling_mode): all_journals

input: url: string – the base URL of the Google Scholar profile
start: integer – the starting index for pagination.
crawling_mode: string – the mode of crawling.

global: status_log: list – global list to store log messages.

output: all_journals: list – a list containing the processed journal data.

```
while attempts < max_attempts:
  Construct page_url with start index and page size = 20
  Log the URL being fetched
  Fetch page content
  If content is missing:
    Log "No content fetched, ending pagination."
    Break loop
  Parse journals from page content
  If no journals found:
    Log "No journals found, ending pagination."
    Break loop
  If crawling_mode == 'detail':
    For each journal in journals do:
      If detail URL exists:
        Fetch additional details from detail page
        Append details to journal record
    Append fetched journals to all_journals
  Else (crawling_mode == 'non-detail'):
    Append basic journal data (first 5 columns) to all_journals
  Log the number of journals fetched
  If less than 20 journals found:
    Log "Less than 20 journals found, assuming end of data."
    Break loop
  Increment start index by 20 for the next page
  Sleep for 1 second to avoid rate limiting
```

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal.768-773
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Algorithm: Data Crawling Algorithm

```
        Reset attempts after a successful fetch
    return all_journals
```

2.3. Data Preprocesing and Storage

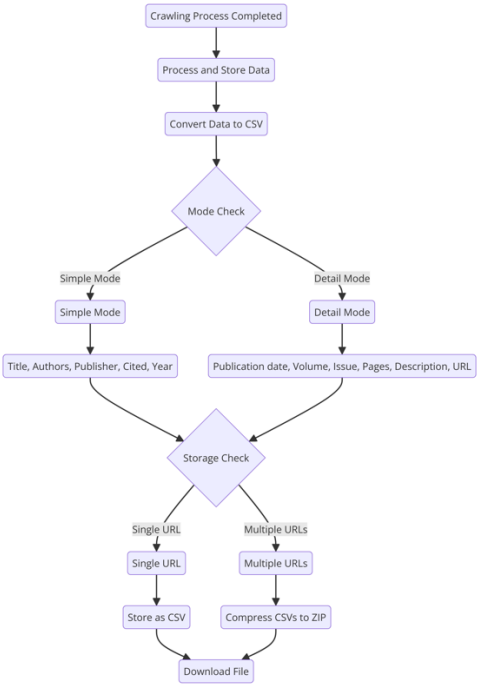

Figure 2. Data Processing and Storage Flowchart

In Figure 2, after the crawling process is completed, the data retrieved from Google Scholar is processed and stored in a format that facilitates further analysis. The collected data is converted into a CSV file, with the column structure tailored to the selected crawling mode.

If the simple crawling mode is used, the resulting CSV file will contain five main columns: Title, Authors, Publisher, Cited, and Year. However, if the detailed mode is selected, the CSV file will include additional columns with more detailed information, such as Publication Date, Volume, Issue, Pages, Description, and URL.

During the storage phase, if only one URL is crawled, the crawling results will be saved as a single CSV file, which the user can download directly. If the crawling process is performed on multiple URLs, each crawling result will be stored in a separate CSV file, and all these CSV files will then be compressed into a single ZIP file to facilitate downloading. The user can download this ZIP file, which contains all the individual CSV files.

2.4. Performance Testing

Performance testing of the crawling tool was conducted to evaluate execution time in two different modes: simple and detailed. This testing involved two author profiles with varying numbers of publications: the first profile contains 41 articles, while the second profile includes 5 articles. In the simple mode, the tool collects only basic data such as Title, Authors, Publisher, Cited, and Year. In contrast, the detailed mode gathers additional information such as Publication Date, Volume, Issue, Pages, Description, and URL, which leads to longer processing times, particularly with larger datasets.

This approach aims to understand how the complexity and size of the dataset affect the tool's performance. The test results showed a significant difference in execution time between the two modes, consistent with previous research findings that the optimization of crawling tools depends on balancing speed with the depth of the data being extracted [18].

2.5. Website

The use of a website as a platform to build the Google Scholar data extraction tool was chosen for its ease of access and flexibility. By using Flask as the framework, this website allows users to access the tool from various devices without the need for additional software installation [19]. The website is designed to provide a simple interface where users can input Google Scholar profile URLs, select the crawling mode (simple or detailed), and download the results in either CSV or ZIP format.

In addition to ease of use, the development of this website also focuses on performance and efficiency in data processing. With the support of modern web technologies, the website can process data quickly and effectively on the server while maintaining an easy-to-use interface [20]. As a result, this tool is reliable for extracting data from Google Scholar and can be accessed by researchers from multiple devices without requiring software installation. Furthermore, the tool is designed to handle varying data sizes efficiently, ensuring that even large datasets are processed in a timely manner, thus making it a robust solution for researchers who need to manage and analyze academic publications comprehensively.

RESULT AND DISCUSSION

Figure 3. Initial View

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal.768-773
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Figure 4. Data Crawling Process with multiple URLs

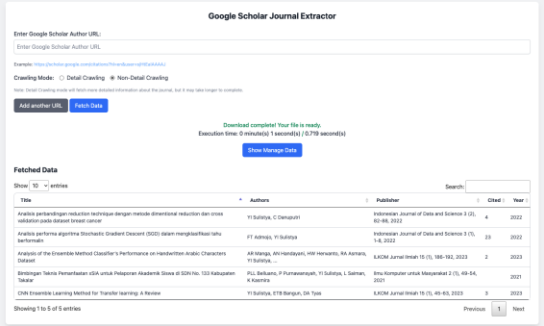

Figure 5. Implementation Testing for Simple Mode

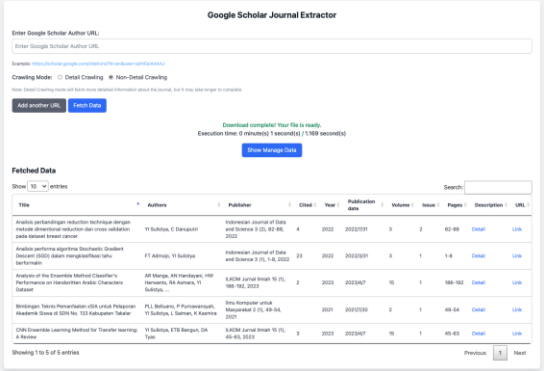

Figure 6. Implementation Testing for Detail Mode

Figure 3 displays the initial page of the website developed for the Google Scholar data extraction tool. On this page, users can enter the URL of a Google Scholar profile and select the crawling mode, either simple or detailed. The page is designed to be straightforward, ensuring that users can easily understand how to operate the tool without any confusion.

Figure 4 illustrates the process where users can input multiple Google Scholar URLs simultaneously. Each URL entered can be easily removed using the delete button, and users can also choose the crawling mode before initiating the data extraction process. This feature highlights the tool's capability to manage multiple URLs at once, which is particularly useful for users who need to gather data from several profiles simultaneously. Additionally, users can monitor the status or progress of the crawling process, providing real-time feedback on the operation.

Figure 5 presents the results of testing the application in simple mode. In this mode, the tool extracts basic information such as the title, authors, publisher, citation count, and publication year of each article. The results are organized into a well-structured table, with the execution time displayed at the top. The simple mode is proven to be fast and efficient, particularly when users only require basic data without the need for detailed analysis.

Figure 6 showcases the results of testing the application in detailed mode. This mode retrieves more comprehensive information from each article, including publication date, volume, issue, pages, description, and URL. The more detailed table provides a complete overview of each article, making it highly valuable for users who require in-depth data. Although the detailed mode requires a longer processing time, the outcome is more thorough and beneficial for those needing more extensive information. This mode is particularly useful for researchers conducting detailed analyses where every aspect of the publication needs to be considered.

Tabel 2. Result Sample of Data Crawling (.csv file)

Title	Authors	Publisher	Cited	Year
Analisis performa algoritma…	FT Admojo, YI Sulistya	Indonesian Journal of Data and Science 3 (1), 1-8, 2022	23	2022
Analisis perbanding an reduction…	YI Sulistya, C Danuputri	Indonesian Journal of Data and Science 3 (2), 82-88, 2022	4	2022
CNN Ensemble Learning…	YI Sulistya, ETB Bangun, DA Tyas	ILKOM Jurnal Ilmiah 15 (1), 45-63, 2023	3	2023
Analysis of the Ensemble...	AR Manga, AN Handayani, HW Herwanto, RA Asmara, YI Sulistya, ...	ILKOM Jurnal Ilmiah 15 (1), 186-192, 2023	2	2023
Bimbingan Teknis Pemanfaatan…	PLL Belluano, P Purnawansyah, YI Sulistya, L Saiman, K Kasmira	Ilmu Komputer untuk Masyarakat 2 (1), 49-54, 2021	0	2021

Table 2 presents a sample of the data crawling results that have been stored in a CSV file format. This table displays essential information extracted from Google Scholar, including the article title,

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal.768-773
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

authors, publisher, citation count, and publication year. The data was collected using the simple mode, which focuses on gathering only the basic information. This dataset provides a general overview of the publications retrieved, such as the article titled "Analisis Performa Algoritma," published in 2022 with 23 citations. Each row in the table represents an individual article that has been successfully crawled from Google Scholar.

After the data fetching process is completed, these results are stored in a CSV file, which users can easily download for further analysis. The CSV format ensures that the data is organized and accessible, allowing for straightforward integration into various data processing tools. This format is particularly useful for researchers who need to analyze or manage large volumes of publication data efficiently. The ability to download the data in CSV format provides flexibility, enabling users to manipulate and explore the data according to their specific research needs.

Figure 7. Execution Time by Profile and Mode

As shown in Figure 7, performance testing of the crawling tool was conducted using two author profiles with differing numbers of publications. For Profile 1, which contains 41 articles, the execution time was 9.054 seconds in simple mode and 71.898 seconds in detailed mode. The longer processing time in detailed mode is attributed to the retrieval of additional information, such as publication date, volume, issue, pages, description, and URL for each article.

For Profile 2, which has only 5 articles, the execution time was 3.283 seconds in simple mode and 11.908 seconds in detailed mode. These results indicate that the detailed mode requires significantly more time, particularly when dealing with larger datasets, whereas the simple mode is faster but provides more limited information. The detailed mode's extended processing time reflects the complexity and depth of the data it collects, making it more suitable for comprehensive analysis. In contrast, the simple mode is ideal for situations where basic data suffices, offering quicker results while sacrificing some detail.

CONCLUSION

Based on the results and discussion, the developed Google Scholar data extraction tool has proven effective in collecting data in both simple and detailed modes. The website simplifies the process for users, allowing them to input one or multiple URLs, select the crawling mode, and obtain the required data. The simple mode enables quick and efficient collection of basic data, with execution times of approximately 9.054 seconds for a profile containing 41 articles and about 3.283 seconds for a profile with 5 articles. In contrast, the detailed mode provides more comprehensive data but requires longer processing times, approximately 71.898 seconds for the profile with 41 articles and about 11.908 seconds for the profile with 5 articles.

Looking ahead, potential future work includes enhancing the tool's capabilities by incorporating features such as automated data analysis, integration with other academic databases, and user-friendly visualization options. Additionally, exploring machine learning algorithms to optimize data extraction and processing times could further improve the tool's efficiency and utility for researchers.

The significant difference in execution times between the simple and detailed modes demonstrates that the tool is highly adaptable and can be tailored to the user's needs, depending on how quickly the data is required and the level of detail desired. Moving forward, there is potential to improve the tool's efficiency, particularly in the detailed mode, to reduce processing time without compromising the completeness of the data collected. This would ensure that the tool continues to provide optimal benefits for researchers who require high-quality data within a more efficient timeframe.

REFRENCES

[1] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 4, pp. 377–383, 2023.

[2] P. Ni, X. Wang, B. Lv, and L. Wu, "GTR: An explainable Graph Topic-aware Recommender for scholarly document," *Electron. Commer. Res. Appl.*, vol. 67, no. 1, pp. 1–10, Sep. 2024.

[3] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PLoS One*, vol. 9, no. 5, pp. 1–6, May 2014.

[4] R. van Dinter, B. Tekinerdogan, and C. Catal, "Automation of systematic literature reviews: A systematic literature review," *Inf. Softw. Technol.*, vol. 136, no. October 2020, p. 106589, 2021.

[5] Q. Liu, R. Yahyapour, H. Liu, and Y. Hu, "A novel combining method of dynamic and static web crawler

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

with parallel computing," *Multimed. Tools Appl.*, vol. 83, no. 21, pp. 60343–60364, Jun. 2024.

[6] C. Bhatt, A. Bisht, R. Chauhan, A. Vishvakarma, M. Kumar, and S. Sharma, "Web Scraping Techniques and Its Applications: A Review," in *Proceedings - 2023 3rd International Conference on Innovative Sustainable Computational Technologies, CISCT 2023*, 2023, pp. 381–394.

[7] N. D. K. A. *et al.*, "Bibliographic dataset of literature for analysing global trends and progress of the machine learning paradigm in space weather research," *Data Br.*, vol. 51, no. 5, pp. 1–8, Dec. 2023.

[8] N. Sateesh, K. Kaur, M. Lakshminarayana, V. Vekariya, H. Patil, and R. Maranan, "Development of a GUI for Automated Classification of Scientific Journal Articles using clustering," in *2024 5th International Conference on Innovative Trends in Information Technology, ICITIIT 2024*, 2024, pp. 1–6.

[9] P. P. Kusumojati and E. Mediawati, "Web-Based Asset Management Information Systems in Higher Education," *Int. J. Business, Law, Educ.*, vol. 5, no. 1, pp. 398–411, 2024.

[10] J. Leskovec, A. Rajaraman, and D. J. Ullman, *Mining of Massive Datasets 3rd*, 3rd ed., vol. 3. 2020.

[11] B. Tóth, L. Berek, L. Gulácsi, M. Péntek, and Z. Zrubka, "Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed," *Syst. Rev.*, vol. 13, no. 1, pp. 1–22, 2024.

[12] L. Yu, Y. Li, Q. Zeng, Y. Sun, Y. Bian, and W. He, "Summary of web crawler technology research," in *Journal of Physics: Conference Series*, 2020, vol. 1449, no. 1, pp. 1–6.

[13] M. Thelwall, "Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals," *J. Informetr.*, vol. 11, no. 4, pp. 1201–1212, Nov. 2017.

[14] S. A. Mohamed, M. A. Mahmoud, M. N. Mahdi, and S. A. Mostafa, "Improving Efficiency and Effectiveness of Robotic Process Automation in Human Resource Management," *Sustain.*, vol. 14, no. 7, 2022.

[15] A. Martín-Martín, R. Costas, T. Van Leeuwen, and E. Delgado López-Cózar, "Evidence of open access of scientific publications in Google Scholar: A large-scale analysis," *J. Informetr.*, vol. 12, no. 3, pp. 819–841, Aug. 2018.

[16] O. Kambli, A. Karande, and H. Kanakia, "EasyChair Preprint H-Index Analysis of Research Paper Using Web Crawling Techniques," in *EasyChair*, 2023, pp. 521–531.

[17] F. M. Javed *et al.*, "An Effective Implementation of Web Crawling Technology to Retrieve Data from the World Wide Web (www)," *Int. J. Sci. Technol. Res.*, vol. 9, no. 1, pp. 1152–1256, 2020.

[18] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained from Web Pages," *IEEE Access*, vol. 8, pp. 61726–61740, 2020.

[19] A. W. Anuar, A. Azmi, N. Kama, H. M. Rusli, N. A. A. Bakar, and N. Mohamed, "Integrating user experience assessment in Re-CRUD console framework development," *Wirel. Networks*, vol. 29, no. 1, pp. 109–127, Jan. 2023.

[20] S. Wan, H. Lin, W. Gan, J. Chen, and P. S. Yu, "Web3: The Next Internet Revolution," *IEEE*, vol. 1, no. 1, pp. 1–11, Mar. 2023.