



# Google Scholar Crawling for Constructing Research Database

M. Miftakul Amin <sup>1</sup>, Adi Sutrisman <sup>2</sup>, Yevi Dwitayanti <sup>3</sup>

<sup>1,2</sup> Department of Computer Engineering, Politeknik Negeri Sriwijaya, Jl. Srijaya Negara Bukit Besar, Palembang, 30139, Indonesia

<sup>3</sup> Department of Computer Accounting, Politeknik Negeri Sriwijaya, Jl. Srijaya Negara Bukit Besar, Palembang, 30139, Indonesia  
miftakul\_a@polsri.ac.id

**Abstract.** This study aimed to compile a research database originating from the Google Scholar dataset using crawling or scraping techniques. The results of the crawl showed that a total of 403 lecturers and researchers at the Sriwijaya State Polytechnic were successfully collected and all 9,511 scientific publications from each lecturer and researcher had been successfully stored in the database. The crawling process was greatly influenced by the validity of the information that was an important part of scientific publication metadata. Similarly, it was also influenced by the number of scientific publications from each lecturer and researcher. The database that has been successfully collected can be displayed in a web application that contains directory information for researchers and lecturers, as well as several parameters for measuring publication performance that has been achieved. This information was useful for management in tertiary institutions to take strategic steps to improve the reputation of tertiary institutions.

**Keywords:** Google Scholar Crawling, Research Database, Web Scraping.

## 1 Introduction

Google Scholar (GS) was introduced in 2004 as a global research database service, where researchers can easily create research profiles on Google Scholar that contain track records of scientific publications, citation information, and H-index information. Various scientific publications can be loaded such as dissertations, articles, papers, reports, books, and so on [1]. The main advantages of this GS are services that were open access [2], and H-index information [3]. According to [4] the relationship between Google Scholar and research institutions and universities was a symbiotic relationship of mutualism. It can be seen that lecturers and researchers from tertiary institutions and research institutions can publish their research results to be loaded for free on Google Scholar.

The productivity of a researcher indicates how much the H-index value has been achieved. Google Scholar has provided researchers with an H-index calculation mechanism automatically from the information stored in the dataset [5]. Google Scholar

was a mediator that contained research community information, which provided free access services and a user-friendly display that made it easy to access [6].

Several investigations related to Google Scholar have been carried out. Research conducted by [7] has developed a model for cleaning misclassified papers in the Google Scholar dataset. This research develops a model to find the wrong entity and clean the wrong item from the dataset. This research has succeeded in eliminating some of the wrong information in grouping paper items in the possession of the author. The Google Scholar dataset has also been used by [8] to develop datasets at the institutional level. Development at the institutional level is very useful for leaders to evaluate research productivity by formulating assessment parameters independently.

Web scraping or crawling is an activity to extract important information from a web page [9]. Web scraping on Google Scholar has been carried out by [10] who managed to collect Google Scholar data in Excel file format and MySQL database. However, there are still limitations to the processed data, namely only one table.

Sriwijaya State Polytechnic currently does not yet have research productivity information services from lecturers and researchers. This opens up opportunities, to develop an application that contains research information services that can compile scientific publications that have been successfully published. The crawling/scraping approach can be used to collect scientific publication data that has been stored in the Google Scholar repository, which can then be developed as a separate publication service.


## 2 Research Method

### 2.1 Dataset


Data was collected from the Google Scholar dataset using the keyword "Sriwijaya State Polytechnic", as can be seen in Figure 1. From the crawling process carried out, data were collected on 403 authors who are researchers and lecturers at Sriwijaya State Polytechnic. Furthermore, each account of this lecturer and researcher will undergo a crawling process to compile all scientific publications that have been produced.

politeknik negeri sriwijaya


---




**Yohandri Bow**  
Politeknik Negeri Sriwijaya  
Email yang diverifikasi di polsri.ac.id  
[Elektrokimia](#) [Energi dan lingkungan](#) [Instrumen dan Kontrol](#)



**Pola Risma**  
Politeknik Negeri Sriwijaya  
Email yang diverifikasi di polsri.ac.id  
[Electrical Engineering](#)



**M. Miftakul Amin**  
Politeknik Negeri Sriwijaya  
Email yang diverifikasi di polsri.ac.id  
[Software Engineering](#) [Artificial Intelligence](#) [Machine Learning](#)



**yurni oktarina**  
Politeknik Negeri Sriwijaya  
Email yang diverifikasi di polsri.ac.id  
[electronic](#) [mechatronic](#)

**Fig. 1.** Initialize the Google Scholar Dataset

## 2.2 Database Design

This research requires persistent data storage in a database management system. MySQL or MariaDB is used to store data that has been successfully obtained. Figure 2 shows the database design consisting of two tables, namely the `gs_author` table to store 403 records of lecturer and researcher information, and the `gs_paper` table to store all scientific publication results from each lecturer and researcher. The `gs_author` and `gs_paper` tables are connected 1:n (one to many) where an author in the `gs_author` table will have a large number of scientific publications stored in `gs_paper`. In the `gs_author` table, there is a primary key in the `user_id` column, while in the `gs_paper` table, there is a foreign key in the `ref_author` column. Even though the column names are different, the contents of the data in these two columns contain the same data.

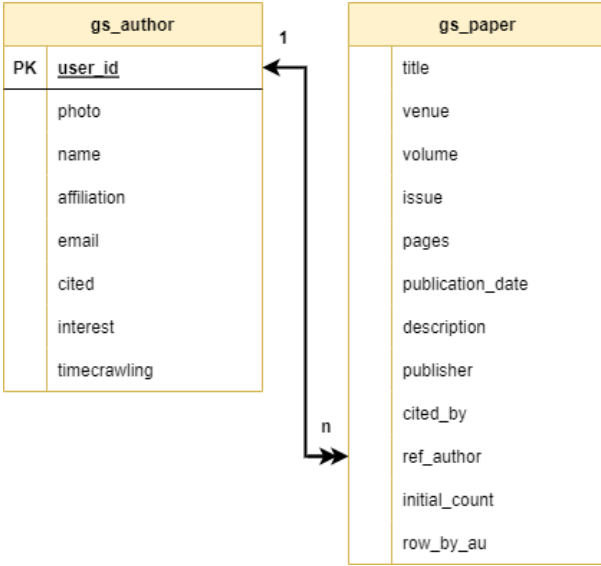


Fig. 2. Table Relationship Design

### 2.3 System Architecture

The Google Scholar Crawling model proposed in this study can be seen in Figure 3. The crawling process is carried out by indexing the Google Scholar repository through the Google Scholar Page by tracing each HTML component from which the data will be retrieved. Furthermore, the crawling process is carried out using the Python programming language and several important libraries such as Selenium, BeautifulSoup, MySQL, Panda Dataframe, and other supporting libraries. This crawling process produces two datasets, namely **gs\_author** and **gs\_paper** which are stored in MySQL.

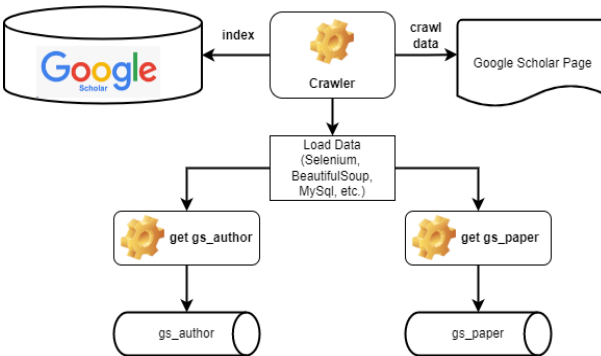
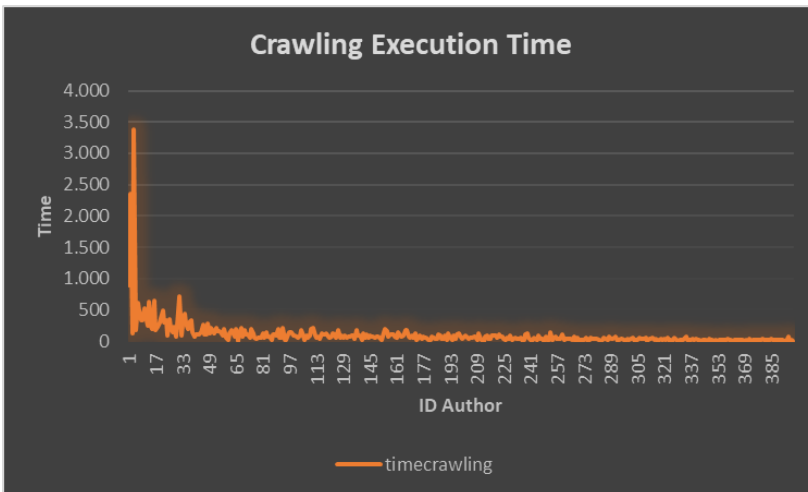


Fig. 3. System Architecture

### 3 Implementation and Results

#### 3.1 Crawling Author

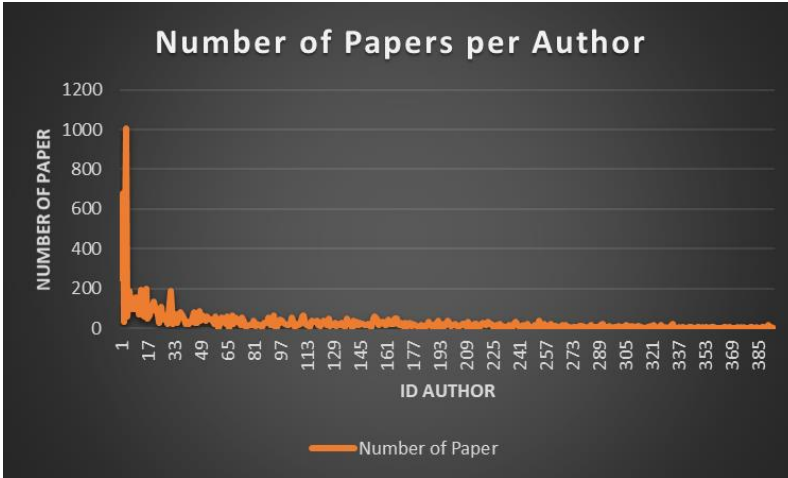
Figure 4 shows the results of crawling from lecturers and researchers at the Sriwijaya State Polytechnic. The results of crawling show that there are 403 authors and the execution time required for crawling. Crawling takes quite a lot of time for the authors in the first order because these authors have many papers and are sorted by Google Scholar on the front page of the Google Scholar web page. The highest crawl time is 3385.2472 seconds, and the smallest is 6.0698 seconds with an average crawl time for each author of 98.5322.



**Fig. 4.** The Time Required for the Paper Crawling Process

#### 3.2 Crawling Paper

The results of crawling the paper are shown in Figure 5. From each author who has successfully crawled in the previous stage, each paper will be crawled using iteration based on the active author. An author has a maximum of 1,003 papers, and at least 1 paper.



**Fig. 5.** Number of Documents for Each Author

#### 4 Conclusion

The results of crawling obtained the number of authors as many as 403, and the number of papers as much as 9,511. The crawling algorithm used in this study has limitations, namely not being able to distinguish crawling results in the form of journals, conferences, books, and other forms of scientific sources. If there is empty metadata, it will be stored as information with a null value in the database. Thus Further research opportunities can improve the performance of the crawling algorithm thus it can be classified according to the category of scientific publications. Further research that can also be developed is to conduct authorization verification. This is looking at the information stored in the Google Scholar dataset, many scientific publications are included in the ownership list of lecturers or researchers who are not authors.

#### References

1. Doğan, G., Şencan, İ., Tonta, Y.: Does dirty data affect google scholar citations?. In: Proc. Assoc. Inf. Sci. Technol., vol. 53, no. 1, pp. 1–4, (2016), doi: 10.1002/pr2.2016.14505301098.
2. Martín-Martín, A., Costas, R., Van Leeuwen, T., Delgado López-Cózar, E.: Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. In: J. Informetr., vol. 12, no. 3, pp. 819–841, (2018), doi: 10.1016/j.joi.2018.06.012.
3. Teixeira da Silva, J. A.: The Google Scholar h-index: useful but burdensome metric. In: Scientometrics, vol. 117, no. 1. Springer International Publishing, pp. 631–635, (2018). doi: 10.1007/s11192-018-2859-7.
4. Oh, K. E., Colón-Aguirre, M.: A comparative study of perceptions and use of google scholar and academic library discovery systems. In: Coll. Res. Libr., vol. 80, no. 6, pp. 876–891, (2019), doi: 10.5860/crl.80.6.876.

5. Van Bevern, R., Komusiewicz, R., Niedermeier, C., Sorge, M., Walsh, T.: H-index manipulation by merging articles: Models, theory, and experiments. In: *Artif. Intell.*, no. August, (2016), doi: 10.1016/j.artint.2016.08.001.
6. Liu, X.: The Google Scholar Experiment: How to Index False Papers and Manipulate Bibliometric Indicators. In: *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. July, pp. 1852–1863, (2013), doi: 10.1002/asi.
7. Hao, S., Xu, Y., Tang, N., Li, G., Feng, J.: Cleaning your wrong google scholar entries. In: *Proceedings - IEEE 34th International Conference on Data Engineering, ICDE 2018*, (2018), pp. 1597–1600. doi: 10.1109/ICDE.2018.00185.
8. Mingers, J., O’Hanley, J. R., Okunola, M.: Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*, vol. 113, no. 3, pp. 1627–1643, (2017), doi: 10.1007/s11192-017-2532-6.
9. Rafsanjani, M. R.: ScrapPaper: A web scrapping method to extract journal information from PubMed and Google Scholar search result using Python. *bioRxiv*, p. 2022.03.08.483427, (2022), [Online]. Available: [https://www.biorxiv.org/content/10.1101/2022.03.08.483427, https://www.biorxiv.org/content/10.1101/2022.03.08.483427v1.abstract](https://www.biorxiv.org/content/10.1101/2022.03.08.483427v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.03.08.483427v1.abstract)
10. Pratiba, D., Abhay, M. S., Dua, A., Shanbhag, G. K., Bhandari, N., Singh, U.: Web Scraping and Data Acquisition Using Google Scholar. in *Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, (2018), pp. 277–281. doi: 10.1109/CSITSS.2018.8768777.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

