Plagiarism Spider: How to Collect Plagiarism from Article

M Zainal Arifin, Suyono, Aji Prasetya W, Harits Ar Rosyid, Daniar W
Universitas Negeri Malang

E-mail: arifin.mzainal@um.ac.id

Abstract. Web crawlers are techniques for gathering information on a website. Web crawlers work by visiting each specified website address then getting and storing all information contained in the website. The Web Crawler technique is commonly used to collect data or information through the internet. The benefit of web crawlers is that the information obtained is more focused so that it makes it easier to do something search. The way the web crawler works is to start the process by providing a list of website addresses to be visited, and every time you visit the address, the crawler will look for another address contained in it and add to the list of previous website addresses. This study provides a proposal about the use of web crawlers in gathering scientific journal information and then the information obtained is reprocessed on a system. Data obtained from the crawler process will be stored in the database and will be processed by the scientific journal management system according to the needs and objectives.

Keyword: web crawler; article; plagiarism; journal

1. Introduction
Internet usage has increased a lot in recent times. Internet is the shared global computing network, that enables global communications between all connected computing devices. The internet provides platform for web services and the World Wide Web (WWW). Web is an information system where documents and other web resources are identified by Uniform Resource Locator (URL), which may be interlinked by hypertext, and accessible over the Internet. In Internet data are highly unstructured which makes it extremely difficult to search and retrieve valuable information.

Web is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child to an adult. There is a spectacular growth in web-based information sources and services. It is estimated that, there is approximately doubling of web pages each year. As the Web grows grander and more diverse, search engines have assumed a central role in the WWW infrastructure as its scale and impact have escalated.

Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. Web crawlers are programs that exploit the graph structure of the web to move from page to page. In their infancy such programs were also called wanderers, robots, spiders, and worms, words that are quite evocative of Web imagery.
Web crawlers or also known as web spiders or web robots are programs that work with certain methods and automatically collect all the information contained in a website. Web crawlers will visit each website address given to him, then absorb and store all information contained in the website. Every time the web
crawler visits a website, he will also list all the links that are on the page he visits to be visited one by one later.

When crawlers find web pages, the next task is to retrieve data from web pages and store them in a storage medium (hard disk). These stored data later can be accessed when the query is related to the data. To achieve the goal of collecting billions of web pages and presenting them in seconds, search engines need a very large and sophisticated data center to manage all this data.

The process of web crawlers in visiting each web document is called web crawling or spidering. The crawling process in a website starts from recording all the URLs of the website, searching one by one, then entering it in the list of pages in the search engine index, so that whenever there is a change on the website, it will be updated automatically. Web crawling is the process of taking a collection of pages from a web for indexing so that it supports search engine performance. One example of a site that implements web crawling is www.webcrawler.com. In addition to site-leading search engines, of course, like Google, Yahoo, Ask, Live, and so on.

Web crawlers are commonly used to make copies in part or in whole web pages that have been visited so that they can be further processed by the index compiler system. Crawlers can also be used to maintain a website, such as validating html code on a web, and crawlers are also used to obtain special data such as collecting e-mail addresses.

Web crawlers are included in the software agent section or better known as the bot program. In general, crawlers start the process by listing a number of website addresses to visit, referred to as seeds. Every time a web page is visited, the crawler will look for another address contained in it and add it to the previous seeds list.

Based on the background above the author is interested to research, design and implementing a search application by using the PHP programming language and MySQL database as a storage medium the data will be used specifically to collect information about articles on scientific documents

2. Scientific Article
A scientific journal is a publication that contains data and information that submits science and technology and is written in accordance with the rules of scientific writing and published periodically. Most paper published in today’s scientific journals use simple structures. With several variations, most paper use the format “IMRaD”: (1) Introduction, (2) Method (experiment, theory, design, model), (3) Result and Discussion, (4) Conclusions [10].

Scientific journals must meet the administrative requirements as follows (1) has an International Standard Serial Number (ISSN), (2) having partners of at least 4 (four) people, (3) Published regularly with a frequency of at least twice a year, except for scientific magazines with specialization scientific coverage with a frequency of once a year, (4) each publication has at least 300 copies, except for scientific magazines that publish electronic journal systems (e-journals) and scientific magazines that implement online systems with the same requirements as printed scientific magazine requirements. (5) contains the main articles each time the issuance amounts to at least 5 (five), besides being added with short communication articles which are limited to a maximum of 3 (three) pieces.
Sources of scientific data and information that are used as the basis for the preparation of scientific papers such as scientific journals are articles containing data and information that advance science and technology and are written according to scientific rules.
3. Crawler

Web crawler are programs that are made to explore the World Wide Web (WWW) systematically and automatically with the aim of collecting data [6]. The structure of the WWW is a graphical structure, which is the link displayed on web pages can be used to open other web pages [8]. This term is also known as the term spidering. The search process is based on the latest data available on the internet. Almost all machines Existing searchers use the concept of crawlers to collect information from the internet as the main component of the search engine [9].

Web crawler is one of the main components of the web search engines. The growth of web crawler is increasing in the same way as the web is growing. A list of URLs is available with the web crawler and each URL is called a seed. Each URL is visited by the web crawler. It identifies the different hyperlinks in the page and adds them to the list of URLs to visit. This list is termed as crawl frontier.

Using a set of rules and policies the URLs in the frontier are visited individually. Different pages from the internet are downloaded by the parser and the generator and stored in the database system of the search engine. The URLs are then placed in the queue and later scheduled by the scheduler and can be accessed one by one by the search engine one by one whenever required. The links and related files which are being searched can be made available whenever required at later time according to the requirements [2].

There are some commonly used Web Crawler Techniques: (1) General Purpose Crawling, (2) Focused Crawling, and (3) Distributed Crawling [7].

![Figure 1. Architecture of Web Crawler](image-url)

Figure 1 shows the architecture of web crawler that has three main components: (1) Crawler frontier, contains the list of unvisited URLs, (2) Page downloader, which download pages from the internet corresponding to the URLs which is retrieved from the crawler frontier, (3) Web repository, receives web pages from a crawler and stores it in the database.
The flow of basic crawler is shown in figure 2.

Web crawler recursively adding URLs to the URLs seed. The working of a web crawler is as follows. First, system initializing the seed URLs then adding it to the frontier. Second, URL from the frontier will be selected to start fetching the web page. Third, the retrieved page will be parsing to extract the URLs. Fourth, adding all the unvisited links to the list of URLs into the frontier. The process from the second step will be repeat until the frontier is empty.
System Design

The crawler system in this study focuses on gathering scientific information on articles from websites that provide or become the source of the scientific article. The main function of the system is shown in figure 3.

This feature is the main function of a system called the crawler engine. Applying the crawler technique starts by inputting the URLs in the web address column and pressing the start button to start the crawler.

Data that has been collected from the crawler process will enter the database and display some of the data on this page, the data obtained has the raw text format.

4. Conclusion

The use of web crawler programs that are built can obtain data or information on scientific articles from the website quite easily. This technique can provide results in the form of a raw text format that can be transformed into other formats as needed.

5. Acknowledgement

This research was supported by Universitas Negeri Malang and Universiti Teknikal Malaysia Melaka.
References

[1] Vandana Shrivastava (2018), “A Methodical Study of Web Crawler”, Journal of Engineering Research and Application, Vol. 8, Issue 11, ISSN: 2248-9622.

[2] Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica (2014), “Web Crawler: Extracting the Web Data”, International Journal of Computer Trends and Technology (IJCTT), Vol 13 No 3, ISSN: 2231-2803.

[3] Ankita Dangre, Vishakha Wankhede, Priyanka Akre, Puja Kolpyakwa (2014), “Design and Implementation of Web Crawler”, International Journal of Computer Science and Information Technologies, Vol 5, ISSN: 0975-9646.

[4] S.Amudha, B.SC., M.SC., M.PHIL. (2017), “Web Crawler for Mining Web Data”, International Research Journal of Engineering and Technology (IRJET), Vol 4, Issue 2, p-ISSN 2395-0072.

[5] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik (2014), “Study of Web Crawler and its Different Types”, IOSR Journal of Computer Engineering (IOSR-JCE), Vol 16, Issue 1, p-ISSN: 2278-8727.

[6] S.S. Dhenakaran and K. Thirugnana Sambanthan (2011), “Web Crawler – An Overview”, International Journal of Computer Science and Communication, Vol 02, No 1, pp. 265-267.

[7] Md. Abu Kausar, V.S.D., Sanjeev Kumar Singh (2013), “Web Crawler – A Review”, International Journal of Computer Applications, Vol 63, No 2, 0975-8887.

[8] Trupti V. Udapure, R.D.K., Rajesh C. Dharmik (2014), “Study of Web Crawler and its Different Types”, IOSR Journal of Computer Engineering, Vol 16, Issue 1, p- ISSN: 2278-8727.

[9] Mironeanu Catalin, A.C. (2017), “An efficient method in pre-processing phase of mining suspicious web crawlers”, International Conference on System Theory, Control and Computing (ICSTCC).

[10] Chris A. Mack (2018), “How to Write a Good Scientific Paper”, SPIE, Bellingham, Washington, 98227-0010 USA.