CEPAT Journal of Computer Engineering: Progress, Application and Technology
Vol. 01, No. 02, November 2022, pp. 56-62
ISSN: 2963-6728, DOI: https://doi.org/10.25124/cepat.v1i02.4879

DarkWeb Crawling using Focused and Classified Algorithm

Putri Rahmasari Yunelfi[1], Agus Setiawan Popalia[2], Fina Fahrani[3], Yudha Purwanto[4], Muhammad Faris Ruriawan[5]
[1,2,3,4,5]Department of Computer Engineering, School of Electrical Engineering, Telkom University, Indonesia

Article Info

Article history:

Received May 25, 2022
Revised June 24, 2022
Accepted August 04, 2022

Keywords:

Focused Crawling
Dark Web
TOR
URL

ABSTRACT

Currently, there are more and more cases of illegal goods transactions and personal data being leaked. Illegal transactions and sales of personal data are usually carried out on the dark web because the web has multiple layers of encryption and anonymity when accessing it. Due to the multi-layered network, it is difficult to access and find out content on the dark web. By using the crawling method as well as TOR you can penetrate and browse content on the dark web. The crawling method on the dark web can use focused crawling which uses a URL approach by viewing URLs that relate to each other on the main URL page on topics that match the keywords entered. In previous studies, the focus has been on crawling on the regular web to get maximum results than general search engines. From the research that has been done previously, the development was carried out by crawling on the dark web to maximize content search, not limited to the surface web. The results of this study are that this focused crawler is able to perform its functionality on the dark web and with high precision results by using 3 main URLs for crawling.

Corresponding Author:

Putri Rahmasari Yunelfi
Department of Computer Engineering
School of Electrical Engineering, Telkom University
Bandung, Indonesia
Email: putriyunelfi@student.telkomuniversity.ac.id

1. INTRODUCTION

The dark web is an intranet whose information is intentionally hidden from surface networks. Accessing them requires specific configuration, methods, and assistive software (TOR) to access these resources. In this layer of the dark web, it is known that there are considerable efforts to maintain user anonymity [1]. The dark web has two activities that users do when accessing, namely legal and illegal activities [2]. An example of abuse of a website with illegal activities is the presence of data that openly provides confidential company, government, and other sensitive information for the misuse of that information. This behavior can have a negative impact on related companies, governments, or even individuals. In addition, illegal activities on the dark web that are often carried out are buying and selling goods/things that are illegal by the state such as drugs, organs, alcohol, etc.

One solution that can be used to overcome the use of the dark web for illegal activities is the use of web crawlers. A web crawler or what is now commonly known as a crawler is software that can surf the web automatically and can extract data from the internet [3]. The web crawler is a crawler feature that, is extensible, can add a lot and expand bandwidth, is robust, can work on the static and dynamic web, is scalable to handle new data and protocols, and quality content can be seen from the selected index, duplicate content can differentiate and remove duplicates executable data submitted in multiple passes, exclude content that is prohibited from crawling, and removes spam from blacklisted URLs with priority below [4]. Web crawlers or spiders can automatically index web pages a crawler database is used to store HTML documents [5]. In web crawler implementation, this can be done by writing socket communication-based crawler, HTTP: protocol-based crawler, PhantomJS based interfaceless browser crawler, and selenium based

interface browser crawler [6]. The web crawler method has been widely circulated at this time, especially for its use on conventional websites. However, the dissemination of sensitive data to the public is not only done on conventional websites. Not many web crawler methods are provided for use on the dark web, considering how to access the dark web is quite difficult and requires a different method from conventional sites.

The crawling method used on the dark web is different because basically, the dark web is a network under the normal internet that has a layer of protection (encryption) to access it. With this multi-layer encryption, not everyone can access the dark web. In addition to multi-layer encryption, the dark web is also commonly used because it is impossible to know the identity of the perpetrators of these illegal transactions due to the anonymity system. Dark web crawling is done through deep crawling of specific web pages on the dark web using TOR proxy [7]. One of the crawling methods that can be used to explore the dark web is the focused crawling method, as it can find links on potentially relevant websites while avoiding irrelevant areas of the site. [8][9]. Focused crawling is a method in which a crawler automatically crawls to a relevant web page. Targeted exploration in this process is carried out at the harvest level obtained from calculations carried out with precision and memory by incorporating priority strategies, learning strategies, evaluation strategies and training strategies to be collected in specific areas before being transferred to local repositories [3][5][10][11]. By crawling focused on the dark web, you can collect various relevant URLs on certain topics from the dark web by classifying them according to the keywords you are looking for.

2. THE COMPREHENSIVE THEORETICAL BASIS

This journal summarizes concepts for browsing the dark web. It explains that the Internet has many layers, and there is an explanation that describes the characteristics and types of operations performed at each layer. The journal also describes general robotic tasks (such as the surface web) and specific tasks on the dark web. To access darknet content using the Darknet Index, extract information that helps security and law enforcement agencies with darknet activity. In the test environment, the crawler is designed to have the ability to simulate user login on the black market, index the entire page, and download the required data from that page [2].

Focused crawling is a method for searching web pages in a focused manner by looking at related keywords. In this focus crawling method, every hyperlink that is related to a given page or keyword will be explored. This study discusses the use of the crawling method to retrieve information sources from the bioinformatics web. In the process, the work steps are taken using the slave-master architecture, where the crawl is considered a slave that can be discarded at any time, and the coordinates are considered as the master. During this research, it is known that using the focus crawl method and the slave-master architecture to get optimal results. In this study, it is also known that the use of this focused crawling method is more effective if the computations are used in parallel [3].

In this study, it is explained how to crawl the dark web and analyze the data that has been collected from the crawling results. When crawling on the dark web, you have some difficulties, for example, getting a URL related to the desired topic, because the combination of URLs on the dark web is usually random. In addition to the difficulty in getting URLs, researchers also find it difficult to explore because basically the dark web is used for illegal transactions, causing web pages to not last long. After getting URLs related to certain topics, crawler results can be collected into a database, and dark web researchers can conduct analysis on the dark web [12].

In this journal, the web crawling method is used to extract data and perform analysis on the web marketplace. In the process, in addition to web crawling, web scraping, scrappy splash, and Power BI are also used. In the process of using the web crawling method, it must be ensured that there is data on the page that is the target of the topic. In addition, in making crawling as much as possible pay attention so that in the search process the page does not get stuck or does not find the desired page. In the search process in the web marketplace, there is a possibility of a block on the web crawling being used. To avoid this possibility, different IP manipulations can be carried out. In the process of conducting the research, web crawling is used to automatically move-related pages in the web marketplace. In this study, splash and scrapy were used to extract the data that had been searched thoroughly. The extracted data can be collected in a dataset with the help of a database [13].

Web crawling often has a lot of problems at the same time, in fact, in some cases, these problems can conflict with each other from updating a website query that changes, to looking for a new website to add, this can give a lot of problems and especially if it has resources This paper describes how the paradigms are directly related to web crawling and search engines, about how to use customizable criteria and then look for the best answers that can deal with these problems [14].

3. METHOD

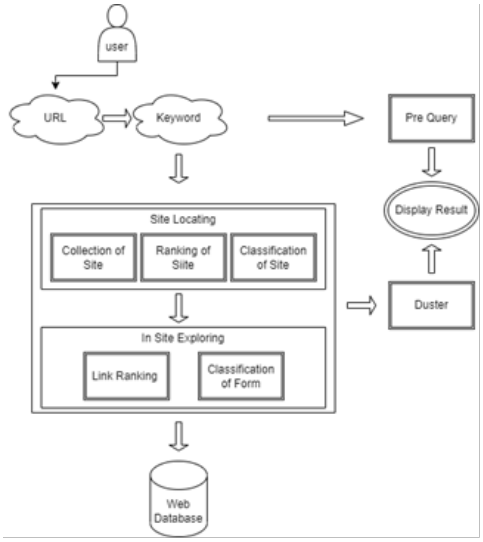

Figure 1. System Architecture

In this system used focused crawling method to explore sites. The crawler will search relevant data on the onion site as quickly and efficiently as possible. This crawler works in the following steps: Site Locating, In-site Exploring, Duster, and Pre-query.

1. User: In this step, the user will provide input a link in the search query. Based on the input given by the user, the crawler will search for relevant content or topic according to the user's needs.
2. Site Locating: During the Site Locating phase, the user will find the most relevant results for a particular topic. It also finds different results in different website formats. The site placement phase consists of three stages, namely: Site Collection, Site Ranking, and Site Classification. This stage will collect all websites both visited and not visited using the reverse search method. After all, sites that have never been visited are collected, a ranking will be given according to relevance and the site will be classified. If the website is relevant, the crawling process will start. Otherwise, the website will be ignored and a new website will be recorded.
3. In-site exploring: At the stage of In-site exploring using 2 stages, namely link ranking and form classification. link ranking serves to give priority to each link that is crawled to get the form. Classification of forms is also carried out to collect and classify the contents of the form to get accurate results
4. Duster: In the crawling results we will find duplicate links. To remove and detect these links can use Duster. Duster is a technique for detecting and eliminating DUST containing duplicate links with similar text from search engines. The use of Duster is also to normalize the data obtained and prevent waste of resources.
5. Pre-Query: A pre-query approach is used to improve the efficiency of the crawling engine. In search by word by word. pre-query will manage history, every time the user will perform a search query it will be checked on the database if there are results in the database it will be given to the user. If the request is not in the database, the data will be searched according to the user's request and the results will be stored in the database.

This program designed to access web pages is a form of web crawling. Search engines for surface web access use the crawl method to get to the page they want to access. Basically, in the process of searching web pages, general crawling on the surface web is done by entering keywords from the information on the page you want to open. Next, the search engine will look for URLs related to the keywords entered and will display a collection of URLs relevant to the keywords entered.

Basically, web crawlers work on the dark web the same way search engines do to crawl the surface web. The method used for crawling on the dark web usually uses a crawl focus to classify the content of web pages, where crawling is done by entering keywords from the information on the page to be accessed as well. However, while crawling the dark web, it is accessed by a special network called TOR to access dark web pages that are encrypted in layers. In crawling the dark web using TOR, it will generate a collection of dark web page URLs that match keywords.

4. RESULTS AND DISCUSSION

3.1. Software Setup

This system is created using the python programming language. This program can run on Windows 8/10/11. Visual Studio Code is used for coding. MySQL is used for making the database. And Glass Fish is used as a web server.

3.2. Performance Measurements

The performance measurement used in this system is Precision. This is used to determine how accurate this system is in retrieving information.
These can be defined below :

$$\text{Precision} = \frac{(Total\ of\ relevant\ URLs\ retrieved)}{(Total\ retrieved\ URLs)} \quad (1)$$

3.3. Crawling Algorithm

Table 1. Algorithm Focused Crawling

```
def crawl (url, keyword)
source = fsources (url)
if(source)
  parse.obj = parse(source)
  url = extract_url(parse.obj)
  return urls
else
  return error
```

3.4. Result

As a result, this paper will show the differences between the existing system and the proposed system. We will compare the accuracy of the two systems. In conducting the experiment, the experiment was carried out 3 times using 3 different input links. For more information about each system results are written in the Table below:

Experiment 1:
Url input: http://nanochanqzaytwlydykbg5nxkgyjxk3zsrctxuoxdmbx5jbh2ydyprid.onion/
Keyword: chan

Table 2. Existing System and Proposed System Result Comparison Url Experiment 1

Result	System Used	
	Existing	Proposed
Total URLs found	142	109
Total relevant URLs	87	93
Precision percentage	**61,26%**	**85,32%**

$$\text{Precision existing} = \frac{87}{142} \times 100 = 61{,}26\%$$

$$\text{Precision proposed} = \frac{93}{109} \times 100 = 85{,}32\%$$

In the table above, it is shown the hat existing crawling method finds more URLs than our system. But, the total relevant URLs found by our system is actually more, and with this number, if we compare system precision between the existing system and ours, the existing system only has 61,26% of precision, this is far less than our system precision result which is 85,32%. The difference in total URLs found between the existing system and ours is due to the difference of focus when crawling is held, where our system, instead of collecting as many as possible URLs first, is actually collecting and calculating the relevance of URLs found sequentially. But, another reason why this experiment has a different result between existing and our system can be caused by internet connection, or how a website server reacts to the crawling process. For some relevant URL results in experiment 1 obtained by our system, it can be seen in table 3.

Table 3. Some Results of Relevant URLs Conducted in Experiment 1

Url Relevant
http://picochanwvqfa2xsrfzlul4x4aqtog2eljll5qnj5iagpbhx2vmfqnid.onion/
http://enxx3byspwsdo446jujc52ucy2pf5urdbhqw3kbsfhlfjwmbpj5smdad.onion/
http://dngtk6iydmpokbyyk3irqznceft3hze6q6rasrqlz46v7pq4klxnl4yd.onion/
http://cct5wy6mzgmft24xzw6zeaf55aaqmo6324gjlsghdhbiw5gdaaf4pkad.onion/
http://wnrgozz3bmm33em4aln3lrbewf3ikxj7fwglqgla2tpdji4znjp7viqd.onion/
http://7sk2kov2xwx6cbc32phynrifegg6pklmzs7luwcggtzrnlsolxxuyfyd.onion/
http://stormwayszuh4juycoy4kwoww5gvcu2c4tdtpkup667pdwe4qenzwayd.onion/
http://xdkriz6cn2avvcr2vks5lvvtmfojz2ohjzj4fhyuka55mvljeso2ztqd.onion/
http://eludemailxhnqzfmxehy3bk5guyhlxbunfyhkcksv4gvx6d3wcf6smad.onion/
http://lainwir3s4y5r7mqm3kurzpljyf77vty2hrrfkps6wm4nnnqzest4lqd.onion/

Table 3 above is some results of the founded and relevant URLs conducted in this experiment 1, with input link http://nanochanqzaytwlydykbg5nxkgyjxk3zsrctxuoxdmbx5jbh2ydyprid.onion/, and searched keyword is “chan”. Table 2 shows some URL results from 93 relevant URLs found by our system.

Experiment 2:
Url input: http://nv3x2jozywh63fkohn5mwp2d73vasusjixn3im3ueof52fmbjsigw6ad.onion/
Keyword: book

Table 4. Existing System and Proposed System Result Comparison Url Experiment 2

Result	System Used	
	Existing	Proposed
Total URLs found	103	78
Total relevant URLs	59	60
Precision percentage	**57,28%**	**76,92%**

$$\text{Precision existing} = \frac{59}{103} \times 100 = 57{,}28\%$$

$$\text{Precision proposed} = \frac{60}{78} \times 100 = 76{,}92\%$$

In table 4, the system crawling result is roughly the same as a result conducted in experiment 1, which is the existing system still has more total URLs found than our system, but in a matter of total relevant URLs, our system is better and its affected the precision percentage which our system has 76,92%, far above the existing system precision percentage which is 57,28%. The difference in total URLs found between the existing and our system is the same as explained in experiment 1.

Table 5. Some Results of Relevant URL Conducted in Experiment 2

Url Relevant
http://archivebyd3rzt3ehjpm4c3bjkyxv3hjleiytnvxcn7x32psn2kxcuid.onion/
http://bible4u2lvhacg4b3to2e2veqpwmrc2c3tjf2wuuqiz332vlwmr4xbad.onion/
http://kx5thpx2olielkihfyo4jgjqfb7zx7wxr3sd4xzt26ochei4m6f7tayd.onion/
http://keybase5wmilwokqirssclfnsqrjdsi7jdir5wy7y7iu3tanwmtp6oid.onion/
http://ciadotgov4sjwlzihbbgxnqg3xiyrg7so2r2o3lt5wz5ypk4sxyjstad.onion/
http://secrdrop5wyphb5x.onion/
http://bcloudwenjxgcxjh6uheyt72a5isimzgg4kv5u74jb2s22y3hzpwh6id.onion/
http://zerobinftagjpeeebbvyzjcqyjpmjvynj5qlexwyxe7l3vqejxnqv5qd.onion/
http://archivecaslytosk.onion/

http://sblib3fk2gryb46d.onion/

Table 5 above is some results of found and relevant URLs conducted in this experiment 2, with input link http://nv3x2jozywh63fkohn5mwp2d73vasusjixn3im3ueof52fmbjsigw6ad.onion/, and searched keyword is “book”. Table 4 shows some URL results from 60 relevant URLs found by our system.
Experiment 3:

Url input: http://p53lf57qovyuvwsc6xnrppyply3vtqm7l6pcobkmyqsiofyeznfu5uqd.onion/
Keyword: news

Table 6. Existing System and Proposed System Result Comparison Url Experiment 3

Result	System Used	
	Existing	Proposed
Total URLs found	196	170
Total relevant URLs	130	148
Precision percentage	**68,42%**	**87,05%**

$$\text{Precision existing} = \frac{130}{196} \times 100 = 68{,}42\%$$

$$\text{Precision proposed} = \frac{148}{170} \times 100 = 87{,}05\%$$

Table 6 above is a result of experiment 3, with a result just like what happened to experiments 1 and 2, an existing system found more URLs than ours, but still, our system has more total relevant URLs found. This affects the precision percentage result, with our system having 87,05% of precision, and the existing system has 68,42%.

Table 7. Some Results of Relevant URL Conducted in Experiment 3

Url Relevant
http://occrpweb4n2vlmih.onion/
http://nytimes3xbfgragh.onion/
http://jokerbuzzhyhl5cl.onion/
http://privacyintyqcroe.onion/
http://xp44cagis447k3lpb4wwhcqukix6cgqokbuys24vmxmbzmaq2gjvc2yd.onion/
http://3qf4wewa5bojmcgr.onion/
http://bfnews3u2ox4m4ty.onion/
http://f3mryj3e2uw2zrv3zv6up6maqosgzn27frz7xodvpl7pkestoyigtkad.onion/
http://dwnewsvdyyiamwnp.onion/
http://27m3p2uv7igmj6kvd4ql3cct5h3sdwrsajovkkndeufumzyfhlfev4qd.onion/

Table 7 above is some results of found and relevant URLs conducted in this experiment 2, with input link http://p53lf57qovyuvwsc6xnrppyply3vtqm7l6pcobkmyqsiofyeznfu5uqd.onion/, and searched keyword is “news”. Table 7 shows some URL results from 148 relevant URLs found by our system.

5. CONCLUSION

This crawler is able to perform its functionality on the dark web and produces high precision. The way this crawler system identifies a given query and starts work by not only collecting every available link but also classifying the results to find the most relevant results and ranking the collected links based on the given query. The system then uses Duster to prevent duplication of all collected links so that the percentage precision represents the true value of and capabilities of this crawler. This paper conducts three experiments on dark web links with different topics. In the first experiment, taking the keyword "chan" on the existing system only has a precision of 61.26%, which is much smaller than our system's precision of 85.32%. In the second experiment using the keyword "book" our system is better and affects the proportion of our system having 76.92%, far above the percentage of the existing precision system which is 57.28%. While the third experiment using the keyword "news" our system has a precision of 87.05%, and the existing system has 68.42%. That way the best percentage level of precision is in our system. But on the use

of crawler focus, we must enter the right keywords, if it does not describe the desired object then the results obtained are maximized.

REFERENCES

[1] S. M. M. Monterrubio, J. E. A. Naranjo, L. I. B. Lopez, and A. L. V. Caraguay, "Black Widow Crawler for TOR network to search for criminal patterns," Proc. - 2021 2nd Int. Conf. Inf. Syst. Softw. Technol. ICI2ST 2021, pp. 108–113, 2021, doi: 10.1109/ICI2ST51859.2021.00023.
[2] B. AlKhatib and R. Basheer, "Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation," J. Digit. Inf. Manag., vol. 17, no. 2, p. 51, 2019, doi: 10.6025/jdim/2019/17/2/51-60.
[3] S. R. Mani Sekhar, G. M. Siddesh, S. S. Manvi, and K. G. Srinivasa, "Optimized focused Web Crawler with Natural Language Processing based relevance measure in bioinformatics web sources," Cybern. Inf. Technol., vol. 19, no. 2, pp. 146–158, 2019, doi: 10.2478/cait-2019-0021.
[4] V. Shrivastava, S. S. J. Subodh, P. G. A. College, and J. National, "OPEN ACCESS A Methodical Study of Web Crawler," VandanaShrivastava J. Eng. Res. Appl., vol. 8, no. 11, pp. 1–8, 2018, doi: 10.9790/9622-0811010108.
[5] K. Velkumar and P. Thendral, "Web Crawler and Web Crawler Algorithms: A Perspective," Int. J. Eng. Adv. Technol., vol. 9, no. 5, pp. 203–205, 2020, doi: 10.35940/ijeat.e9362.069520.
[6] T. Fang, T. Han, C. Zhang, and Y. J. Yao, "Research and construction of the online pesticide information center and discovery platform based on web crawler," Procedia Comput. Sci., vol. 166, pp. 9–14, 2020, doi: 10.1016/j.procs.2020.02.004.
[7] P. Koloveas, T. Chantzios, C. Tryfonopoulos, and S. Skiadopoulos, "A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence," Proc. - 2019 IEEE World Congr. Serv. Serv. 2019, vol. 2642–939X, no. i, pp. 3–8, 2019, doi: 10.1109/SERVICES.2019.00016.
[8] M. Kumar, A. Bindal, R. Gautam, and R. Bhatia, "Keyword query based focused Web crawler," Procedia Comput. Sci., vol. 125, pp. 584–590, 2018, doi: 10.1016/j.procs.2017.12.075.
[9] A. Khazaie, N. B. Seghouani, and F. Bugiotti, Smart Crawling: A New Approach toward Focus Crawling from Twitter, vol. 1, no. 1. Association for Computing Machinery, 2021. [Online]. Available: http://arxiv.org/abs/2110.06022
[10] P. Mishra and A. Khurana, "Accuracy Crawler: An Accurate Crawler for Deep Web Data Extraction," 2018 Int. Conf. Control. Power, Commun. Comput. Technol. ICCPCCT 2018, pp. 25–29, 2018, doi: 10.1109/ICCPCCT.2018.8574286.
[11] D. S. Santoso and R. V. H. Ginardi, "Kompresi Multilevel Pada Metaheuristic Focused Web Crawler," JUTI J. Ilm. Teknol. Inf., vol. 17, no. 1, p. 52, 2019, doi: 10.12962/j24068535.v17i1.a785.
[12] Y. Yang, G. Zhu, H. Yu, and L. Yang, "Crawling and analysis of dark network data," ACM Int. Conf. Proceeding Ser., pp. 116–120, 2020, doi: 10.1145/3379247.3379272.
[13] I. N. Husada, E. H. Fernando, H. Sagala, A. E. Budiman, and H. Toba, "Ekstraksi dan Analisis Produk di Marketplace Secara Otomatis dengan Memanfaatkan Teknologi Web Crawling," J. Tek. Inform. dan Sist. Inf., vol. 5, no. 3, pp. 350–359, 2020, doi: 10.28932/jutisi.v5i3.1977.
[14] A. Gupta and G. N. Campus, "Web Crawling Model and Architecture," no. May, 2021.