Automation Distributed Cloud Based Crawler

Lanang Prismana*

Department of Informatics Engineering, Faculty of Engineering, Universitas Negeri Surabaya
**Corresponding author. Email: lanangprismana@unesa.ac.id*

ABSTRACT
Information is very important data and is needed in various needs. Online news is one type of site that ranks in the top 10 most visited by internet users in Indonesia. Online news sites publish articles to the internet every minute. An online news corpus is necessary for information processing. Retrieval of online news corpus in general has obstacles such as large resource requirements, delays due to excessive access restrictions categorized as bots / spam, thus affecting the speed of retrieval of information from online news. To overcome this problem, it is necessary to develop a framework to improve performance in the creation of an online news corpus. In this study, a framework was developed in creating an online news corpus based on distributed cloud based crawler automation using the MCDM method. The process of self-optimization of cloud tasks in research uses a topsis approach method with alternative data as objects to be assessed, then the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternatives. This framework divides the crawler system and information extraction into several sub-systems. The first stage developed a distributed crawler system, a mechanism for distributing work using a node selection mechanism. The second stage is to develop an information extraction system using a combination of pattern based and node density. The third stage developed automated node management. The contribution of this research is the automation of distributed cloud-based crawler framework which has not been done by previous researchers. This framework activates nodes according to the priority of existing work so that it can speed up the process of retrieving information by using small resources. The performance of this framework will be tested for the accuracy of the extraction results and the average time required. The stages carried out in this research start from URL collection, URL filtering, scheduling, accessing URLs and data extraction. This research focused on the automation of distributed cloud-based crawlers.

Keywords: *crawler, automation, distributed systems, fog-cloud, distributed cloud.*

1. INTRODUCTION

The Internet is a very large network and forms a network of computers that are interconnected throughout the world. With the internet, all data can be collected quite easily. Most of the data on the internet is information that continues to grow every day. Indonesia is one of the countries with the largest population of internet users in the world. According to a report by We Are Social (Hootsuite), there are 204.7 million internet users in the country as of January 2022. That number edged up 1.03% compared to the previous year. In January 2021, the number of internet users in Indonesia was recorded at 202.6 million. The trend in the number of internet users in Indonesia has continued to increase in the last five years (Annur, 2022). Meanwhile, based on data released by Internet world stats in June 2021, it has reached 212 million internet users in Indonesia and is in third place with the most internet usage in Asia. Internet users in Indonesia in getting information prefer to access through online news sites. Based on Similarweb data, news sites rank 2 of the 10 most frequently accessed sites. While based on Alexa, news sites rank 6 of the 10 most frequently accessed sites. Thus making online news sites become sites that are often visited by Indonesians in search of information.

In 2018, the Minister of Communication and Information Technology estimated that currently in Indonesia there are 43 thousand online news sites. However, the number of online media that have been verified by the Press Council is no more than 100 media. Online news is a source of information that is often accessed and used as a reference, so that information in online news becomes a source of big data, because information on online news sites in the form of news articles every minute is always updated and widely published. Big Data can be defined as information assets characterized by high volume, speed, and variety and require certain analytical methods and technologies to convert them into value (Chen, et.al., 2012). There is more information than the person doing the analysis, and this is a potential problem where a lot of data can be ignored (Ma. Tian J., et.al, 2020) [1].

The explosive development of the internet makes it difficult for a person to find information that suits what is desired. In related posts, someone often gets information that is not in accordance with what is desired. The amount of data on the Internet is becoming ever larger and website technology is constantly changing to be a major challenge to

Y. A. Yusran et al. (eds.), *Proceedings of the 2023 Brawijaya International Conference (BIC 2023)*, Advances in Economics, Business and Management Research 294,
https://doi.org/10.2991/978-94-6463-525-6_3

deal with big and complex data on the global Internet. The amount of information accompanied by advertisements displayed on each online news page makes users less comfortable, so that the user's goal in finding the information needed is not focused on the online news content. This is a problem that is often faced by online news users and a system is needed in taking the essence of information from online news. In collecting this information, a program is needed that can obtain the information needed in accordance with the keywords entered, so that the information produced is appropriate and appropriate. The program is built using the concept of a crawler that browses one or more web pages on the internet. In addition, it is also necessary to design the user interface of the program so that it is easily understood by users.

Indexing news articles has many benefits, including as a track record of digital publications, as a source of hoax validation data, specific sources of information, and others. Article indexing includes inserting new articles, updating article data, and deleting articles. The indexing task is done through several stages, search is the first stage of indexing. The search is done by a machine with the name web searcher (web crawler). A web searcher is a machine that visits and collects documents or articles on online sites (Kumar et.al, 2016). The search engine checks the database of articles related to actions to be taken from search results. As the number of online news increases, the role of web crawlers is very important to download and display website information in a short time according to user wishes (Raganan et.al, 2020).

In addition to problems in displaying the main content of online news, other problems related to searching online news articles in general require large resources and take a long time. Online news sites generally provide restrictions for search engines to check the content of their sites. In addition to restrictions from the site, the server resources used are also impacted. These resources include CPU, memory, and bandwidth. In addition, problems also occur when the database and crawler resources are still centralized, so that a long search process will burden system performance which can result in system failure (Kim, et al. 2020).

With a centralized computer system, it takes a large resource to search for articles. Searcher distribution is a solution so that required resources can be minimized. With a distributed system is also a solution to limit the amount of access imposed by news article providers. One distributed system that is a solution is the fog-cloud system. The process of searching for news articles is always correlated with the classification process or the inference process. Both of these processes require several predecessor processes including hyphen removal, common word removal, and tokenization. This precursor process can be carried out at distribution points so that it will relieve the resources of the central computer system.

The contribution of researchers in the research to be carried out is to build an automated distributed cloud-based crawler for searching online news articles so that the process of searching articles on the internet is more efficient because of the division of tasks in each sub-system that is interconnected and does not burden the performance of the central server, and reduces the risk of system failure. In this study, a framework was developed in creating an online news corpus based on distributed cloud based crawler automation using the MCDM method. The process of self-optimization of cloud tasks in research uses a topsis approach method with alternative data as objects to be assessed, then the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternatives. This research aims to produce a distributed cloud-based crawler automation system for online news search that is organized through the work scheduling system of each distribution point (sub-system), so that the online news search process becomes more efficient.

2. RELATED WORK

2.1 Search Engine

Search engines (search engines) are facilities used to explore various data, information, and knowledge that exist on the internet. A search engine is a program that can be accessed via the internet that serves to help computer users in searching for various things they want to know [2]. The Americanican Heritage Dictionary defines a search engine as a software program that searches, captures, and displays information from databases. Searches by Search engines are carried out in a database that stores the text of each page. Text from the page by page is saved into the database server. When performing a search, search engines will search for copies of pages stored in a database containing copies of pages at the time they were last visited. When the link provided is clicked, the address will be given from the search engine server.

2.2 ***Focus Crawler***

Summary of web crawler technology research (Linxuan Yu et al., 2020) this article clearly describes the types of web crawlers that are developing today. Based on its characteristics, there are distributed, scalability, performance and efficiency, quality, freshness, and extensibility. Based on the classification, there are general, focus, incremental, distributed, parallel, and IoT. iCrawl: Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling (Gossen G., Demidova E., and Risse T., 2015), this article describes how to monitor, collect, and analyze the novelty of online articles on the topic of the Ebola virus or the crisis in Ukraine. This article discusses how to overcome the problem of article novelty by interlinking between social media and the web. Both articles above describe that focus crawl prioritizes certain criteria. These criteria can vary, such as topic, time, or object. In general, focus crawls learn their criteria based on initial seeds. Many variants have evolved by applying multiple methods as selection criteria to the focus crawl [3].

2.3 Automation

Automation according to the KBBI digital dictionary is the waiting for human power with machine power that automatically performs and manages work so that it no longer requires human supervision. Automation is done to achieve the goal of making it easier for humans to perform complex and detailed tasks.

2.4 MCDC

Multi Criteria Decision Making (MCDM) is a method of decision making to determine the best alternative from a number of alternatives based on certain criteria. Criteria are usually measures or rules or standards used in decision making. In general, it can be said that MCDM selects the best alternative from a number of alternatives (Kusumadewi, et al. 2006).

2.5 AHP

The striking difference between the AHP method and other decision-making methods lies in the type of input. Existing methods generally use quantitative inputs. Automatically the method can only process quantitative things as well. The AHP method uses human perception that is considered 'expert' as its main input, the 'expert' criterion here does not mean that the person must be genius, smart, have a doctorate degree and so on, but rather refers to people who better understand the problem posed, feel the consequences of a problem or have an interest in the problem. Because it uses qualitative input (human perception), AHP can process quantitative things in addition to qualitative things.

The working principle of AHP is the simplification of a complex problem that is not structured, strategic, and dynamic into its parts, and arranges in a hierarchy. Then the importance of each variable is subjectively given a numerical value about the relative importance of that variable compared to other variables. From these various considerations, synthesis is then carried out to determine variables that have high priority and play a role in influencing the results of the system (Marimin, 2004).

2.6 Topsis (Multiple Criteria Decision Making)

TOPSIS uses the principle that the chosen alternative should have the closest distance from the positive ideal solution and the furthest from the negative ideal solution from a geometric point of view by using Euclidean distances to determine the relative proximity of an alternative to the optimal solution. A positive ideal solution is defined as the sum of all the best achievable values for each attribute, while the negative ideal solution consists of all the worst values achieved for each attribute. TOPSIS considers both, distance to positive ideal solution and distance to negative ideal solution by taking proximity relative to positive ideal solution. Based on comparisons to relative distances, alternative priority arrangements can be achieved.

2.7 Web Crawler

Web crawlers, often known as web robots or web spiders, are software applications that may download web pages more than once and automatically extract data or URLs as desired by users. In general, web crawlers are of two types, namely general web crawlers (GWC) and distributed web crawlers (DWC) depending on the number and operation of machines. DWC can then be categorized as multi-threaded or client-server. Figure 1 shows the process of web crawling data collection, which is repeated until all the data is collected [4].

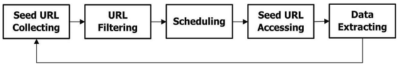

Figure 1 Web crawling process

2.8 Distributed Web Crawling

The distributed web crawler is divided into two parts, the first is multi-threaded (MT DWC), which makes various data threads make one in its entirety from either one machine, or server-client (SC-DWC), where several machines gather data concurrently. The second part is MT-DWC which has an overview like with SC DWC in one machine. The project load is divided into two, namely as follows: number one that works on the seed URL and one that makes one data taken from the website. The advantage of MT-DWC is that it is more economical to make new machines. Distributed web crawlers with a wide variety of machines use server-client designs to shorten data retrieval times [3].

2.9 Fog Computing

Fog computing introduces a layer between edge devices and the cloud. This layer relies on a group of small computing servers that are near the edge device and not necessarily on the device itself. Servers are connected and cloud servers are centralized, allowing for an intelligent flow of information [10]. These small units work together to handle data pre-processing, short-term storage, and rule-based real-time monitoring. Fog computing architecture reduces the amount of data transported through the system and improves overall efficiency [11]. An overview of the fog computing architecture is shown in figure 2.

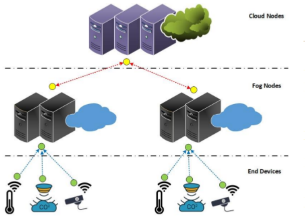

Figure 2 Overview of fog computing architecture

2.10 Cloud Computing

Cloud computing is a network of several devices, computers, and servers that are connected via the Internet [12]. Cloud computing requires storage and access to enter data and programs via the internet from a computer (hardware). Users who use cloud computing do not have a structural infrastructure. This cloud computing realizes itself as a derivative of several other areas of computing. In the cloud market, there are 3 related parties in it. The three parties are as follows, End-user, Business management and Cloud service provider [13].

3. METHODOLOGY

In this case, researchers collect research objects and data on several online news sites in Indonesia, including: detik.com, tribunnews.com, kompas.com, liputan6.com, merdeka.com, kapanlagi.com, okezone.com, tempo.co, viva.co.id, suara.com, sidonews.com, and jawapos.com. The research design used in the development of the distributed news crawler is depicted in figure 6 below:

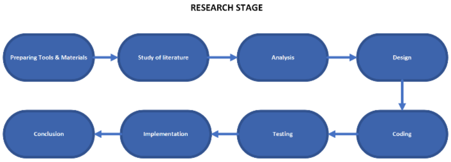

Figure 3 Research stage

At the development stage, distributed cloud-based crawler automation will be carried out using the MCDM method. The process of self-optimization of cloud tasks in research uses the Topsis approach method with alternative data as objects to be assessed, which is divided into 3 alternative data, namely edge 1, edge 2 and edge 3. Furthermore, the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternative.

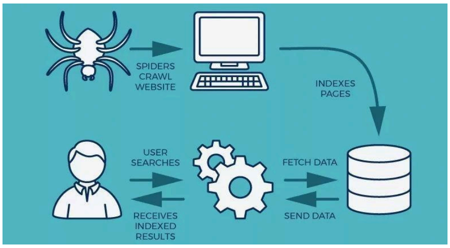

Figure 4 How web crawlers work

4. RESULT

The research flow of the distributed web crawler compiled in this study is as shown in figure 7 below:

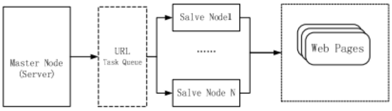

Figure 5 Distributed web crawler system structure

The workflow of this study is as follows: Programming by initializing crawler data, providing part of the cluster. Job assignment of slave node job tracker by master node. Multi-threaded opening by the slave node, and after the task is received will be carried out download the web page. Parsing web pages and storing Web material URLs. Adding URLs to the list that the crawler will assign after filtering and other operations. The division of tasks by the master nodes and see the status of each node. The section node monitors the communication performed with the works on these tasks as a master node via TaskTracker. JobTracker has the task of holding and managing all system communications, as well as allocating tasks. TaskTracker has the authority to work on the Map section and take on Tasks.

Combining with a combination of the system's business process solving, the distributed web crawler has five modules, namely the the download page module, the page parsing module, the URL task allocation module, the parameter configuration module, the data update module, and the indexing module (if needed). Figure 6 of the following is displayed:

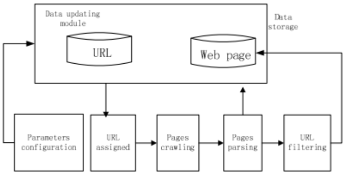

Figure 6 Distributed web crawler system structure

The stages of research conducted in this study are as follows:

4.1 Planning Phase

In this stage there are several important points that need to be made in the process of automation of distributed cloud-based crawlers, including feasibility studies, namely making feasibility studies for the system to be created, scope (scope) which determines the scope limits of the system to be built, namely automation of distributed cloud-based crawlers.

4.2 Analysis Phase

At this stage, researchers analyze the structure and flow of distributed cloud-based crawler automation. The results of the analysis of the structure and flow of the system will be described in the form of business processes. All analysis results will be documented and used as guidelines when designing features.

4.3 Design Stage

The design process transforms needs into characteristic forms that the software understands before the system development process begins. This design must be well documented and part of the software configuration.

4.4 Development Stage

Automation using the MCDM method is further developed. The process of self-optimization of cloud tasks in research uses the Topsis approach method with alternative data as objects to be assessed, which is divided into 3 alternative data, namely edge 1, edge 2 and edge 3. Furthermore, the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternative. There are three rules that web crawlers must consider: relevance and how important a page is, reviewing the page, and robot requests.txt.

The workflow of this system is, when there is a command from the user in searching for information, the activity will be captured by HQ, then from HQ will distribute to each hospital under it. The process flow is as seen in Figure 7.

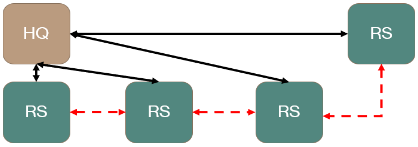

Figure 7 Main System

In the main HQ system, there are also several other sub-systems that work together in conducting information searching, while the process flow in HQ in detail can be seen in Figure 8 below:

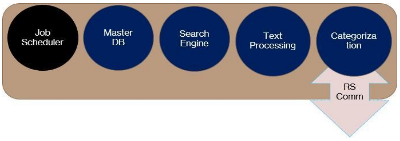

Figure 8 Detail Sistem HQ

RS is part of the HQ system under it, where the task of RS is to receive commands forwarded by HQ sourced from user activity, as for the workflow as seen in Figure 9 below:

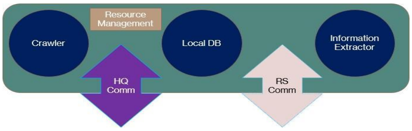

Figure 9 Desain Remote Server

The self-optimization process of the fog cloud task in research uses a Topsis approach with alternative data as the object to be assessed, which is divided into 3 alternative data, namely edge 1, edge 2 and edge 3. Furthermore, this study will use 4 criteria CPU, Memory, CR Load, TE Load as shown in the following figure 10:

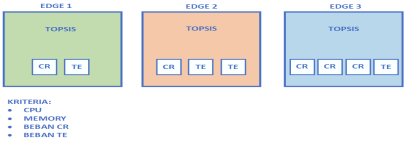

Figure 10 Self-optimization process tasks

After the alternative data and criteria are determined, an alternative value will be sought which will be used to provide an assessment of the alternatives on each criterion. Furthermore, if all data is prepared (alternative data, criteria data, and alternative value data), it will proceed to the calculation process. The following are the steps for calculating the topsis method, namely Normalization, Weighted Normalization, Ideal and Total Solution Matrix, so that you will get the best alternative.

Furthermore, for the task scheduling process, the selection of edge nodes in this study will apply the AHP method, with alternative data as objects to be assessed, namely: URL, TASK and SCHEDULE as shown in figure 11 below:

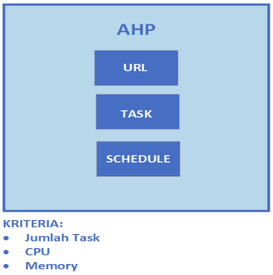

Figure 11 Edge node selection task scheduling process

Then for the criteria data that will be used as a reference / basis for the assessment, namely: Number of tasks, CPU and Memory. After the alternative data and criteria are determined, a comparison value will be sought which will be used to provide an assessment of the alternatives on each criterion. In AHP the comparison value is given between 1 to 9 according to Saaty's theory. In the AHP method, we do 2 comparisons, namely:

1. Comparison between criteria

   Each criterion will be compared against all criteria (including the criteria themselves)

2. Comparison between alternatives

   The concept of assigning value to alternatives is almost the same as the criteria. The advantage is that in the alternative we do comparisons for all criteria.

Furthermore, if all data is prepared (alternative data, criteria data, and comparison value data), it will proceed to the AHP calculation process to get the best alternative. Here are the steps for calculating the SPK AHP method:

1. Calculation of Priority Weight Criteria

   - Search the total row
   - Normalize matrix &; priority weights
   - Find matrix consistency

2. Calculation of Alternative Priority Weights

To find the priority weight of the criteria in the alternative is done as many as the number of criteria, the steps are the same as finding the priority weight of the criteria.

3. Ranking

Based on the priority weight of the criteria and alternative weights, a ranking table can be compiled. Furthermore, to find the total value by multiplying the priority weight of the criterion by each row of the alternative priority weight matrix, the best alternative will be obtained

4.4 Testing Phase

Testing after the design and demonstration / simulation is obtained, testing is carried out on the system that has been developed, to find out the extent of the success rate of system implementation, and find out the shortcomings of the system which can later be updated to reduce system errors. Trials are conducted per section (crawl, content extraction, classification.

4.5 Maintenance

Operations are carried out on the test results obtained. The analysis aims to provide an overview of the condition of the application and input on the direction of further development. Furthermore, maintenance will be carried out on the distributed cloud-based crawler automation system that has been successfully developed. It is hoped that the results of the research can contribute to overcoming existing problems. In this study, the divisional technique involves combining the two theories to get rid of the accumulation of both links and material. The methodology used in this study provides unmistakable proof that detailed web pages are more likely to merge into sites linked to the domain they contain than randomly picked pages.

AUTHORS' CONTRIBUTIONS

In this study, a framework was developed in creating an online news corpus based on distributed cloud based crawler automation using the MCDM method. The process of self-optimization of cloud tasks in research uses a topsis approach method with alternative data as objects to be assessed, then the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternative

ACKNOWLEDGMENTS

In this study, a framework was developed in creating an online news corpus based on distributed cloud based crawler automation using the MCDM method. The process of self-optimization of cloud tasks in research uses a topsis approach method with alternative data as objects to be assessed, then the task scheduling process of selecting edge nodes in this study will apply the AHP method to get the best alternatives. This framework divides the crawler system and information extraction into several sub-systems. The first stage developed a distributed crawler system, a mechanism for distributing work using a node selection mechanism. The second stage is developed information extraction system using a combination of pattern based and node density. The third stage developed automated node management. The contribution of this research is the automation of distributed cloud-based crawler framework which has not been done by previous researchers. This framework activates nodes according to the priority of existing work so that it can speed up the process of retrieving information by using small resources. The performance of this framework will be tested for the accuracy of the extraction results and the average time required. The stages carried out in this research start from URL collection, URL filtering, scheduling, accessing URLs and data extraction. This research focused on the automation of distributed cloud-based crawlers.

REFERENCES

[1] Achsan, H. T. Y., & Wibowo, W. C. (2014). A fast distributed focused-web crawling. *Procedia Engineering*, *69*, 492-499.

[2] Lawrence, S., & Giles, C. L. (1998). Searching the world wide web. *Science*, *280*(5360), 98-100.

[3] Bal, S. K., & Geetha, G. (2016, February). Smart distributed web crawler. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1-5). IEEE.

[4] Ridho, F. (2020). *Rancang bangun aplikasi web crawling untuk mencari harga barang termurah dari berbagai e-marketplace studi kasus: tokopedia, bukalapak, shopee* (Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta).

[5] Pu, Q. (2016, August). The Design and Implementation of a High-Efficiency Distributed Web Crawler. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 100-104). IEEE.

[6] Yong-Young Kim. (2019). Implementation of hybrid P2P networking distributed web crawler using AWS for smart work news big data, Peer-to-Peer Networking and Applications https://doi.org/10.1007/s12083-019-00841-0.

[7] Shokouhi, M., Chubak, P., & Raeesy, Z. (2005, April). Enhancing focused crawling with genetic algorithms. In *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II* (Vol. 2, pp. 503-508). IEEE.

[8] S. Chakrabarti, S., Van den Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer networks*, *31*(11-16), 1623-1640.

[9] Ye, Y., Ma, F., Lu, Y., Chiu, M., & Huang, J. (2004). iSurfer: A focused web crawler based on incremental learning from positive samples. In *Advanced Web Technologies and Applications: 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April 14-17, 2004. Proceedings 6* (pp. 122-134). Springer Berlin Heidelberg.

[10] A. lakhan, Mazin Abed Mohammed, Dheyaa Ahmed Ibrahim et al., Bio-inspired robotics enabled schemes in blockchain-fog-cloud assisted IoMT environment, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2021.11.009

[11] Zainudin, Ahmad, et al. (2021). Implementation of Fog Computing in Smart Home Applications Based on the Internet of Things Cess (Journal of Computer Engineering System and Science) p-ISSN :2502-7131, Vol. 6 No. 1 January 2021

[12] Wu, M., & Lai, J. (2010, December). The Research and Implementation of parallel web crawler in cluster. In *2010 International Conference on Computational and Information Sciences* (pp. 704-708). IEEE.

[13] Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends® in Information Retrieval*, *4*(3), 175-246.

[14] Hussein, M. K., & Mousa, M. H. (2020). Efficient task offloading for IoT-based applications in fog computing using ant colony optimization. *IEEE Access*, *8*, 37191-37201.