Journal of Science, Computing and Engineering Research (JSCER)
Volume-7, Issue-9, September 2024.
DOI: *https://doi.org/10.46379/jscer.2024.0709*
ISSN: 2708-1079

Content Based Filtering Model Based On Segmentation of Webpage Personalization Issue

Riya Sharma and Makesh Singh

Assistant Professor,Raksha Shakti University, Ahmedabad, India.

Article Information

Received : 02 Sept 2024
Revised : 06 Sept 2024
Accepted : 18 Sept 2024
Published : 22 Sept 2024

Corresponding Author:
Riya Sharma

Abstract— In the view of massive content explosion in World Wide Web through diverse sources, it has become mandatory to have content filtering tools. The filtering of contents of the web pages holds greater significance in cases of access by minor-age people. The traditional web page blocking systems goes by the Boolean methodology of either displaying the full page or blocking it completely. With the increased dynamism in the web pages, it has become a common phenomenon that different portions of the web page holds different types of content at different time instances. This paper proposes a model to block the contents at a fine-grained level i.e. instead of completely blocking the page it would be efficient to block only those segments which holds the contents to be blocked. The advantages of this method over the traditional methods are fine-graining level of blocking and automatic identification of portions of the page to be blocked. The experiments conducted on the proposed model indicate 88% of accuracy in filtering out the segments.

Keywords: ***Content Filtering, Segmentation, Web Page Blocking***

Citation: Riya Sharma and Makesh Singh, "Content Based Filtering Model Based On Segmentation of Webpage Personalization Issue", Journal of Science, Computing and Engineering Research, 7(9), September 2024.

I. INTRODUCTION

In the view of massive content explosion in World Wide Web through diverse sources, it has become mandatory to have content filtering tools. The filtering of contents of the web pages holds greater significance in cases of access by minor-age people. The traditional web page blocking systems goes by the Boolean methodology of either displaying the full page or blocking it completely. With the increased dynamism in the web pages, it has become a common phenomenon that different portions of the web page holds different types of content at different time instances. This paper proposes a model to block the contents at a fine-grained level i.e. instead of completely blocking the page it would be efficient to block only those segments which holds the contents to be blocked. The advantages of this method over the traditional methods are fine-graining level of blocking and automatic identification of portions of the page to be blocked. The experiments conducted on the proposed model indicate 88% of accuracy in filtering out the segments.

• Proposing a model for web page content filtering based on segmentation.

• Incorporation of personalization in the proposed model to enhance the web content filtering process.

condition so a large current required enabling fast regeneration in the circuit.

II. RELATED WORKS

This section would highlight the related works that have been carried out in this domain. The proposed model incorporates the following two major fields of study:

• The Web Content Filtering

• Web Page Segmentation

2.1 The Web Content Filtering

Content Filtering Systems for web pages is an active research topic primarily due to following reasons: It protects users (especially minor-age people) from unwanted content; the resources on the network can be saved from unwanted usage like playing network games in an office network etc. There exist many approaches to Content Filtering Systems. Some of them are as listed below:

• Rating Systems

• Black Listing / White Listing

• Keyword blocking

• In Rating Systems users are asked to rate a web site for its content. This rating would be used as a tool for filtering [1]. The black listing / white listing maintains a set of URLs manually prepared for filtering.

The problem with this approach is the scalability. There exist many tools available to perform content filtering using above specified methods [2], [3], [4]. The text classification based approach is explored in [5], [6]. The approach that has been chosen to facilitate filtering in this paper is a variation of keyword based blocking method.

2.2 Web Page Segmentation

Web page segmentation is an active research topic in the information retrieval domain in which a wide range of experiments are conducted. Web page segmentation is the process of dividing a web page into smaller units based on various criteria. The following are four basic types of web page segmentation method:

• Fixed length page segmentation

• DOM based page segmentation

• Vision based page segmentation

• Combined / Hybrid method A comparative study among all these four types of segmentation is illustrated in [7].

Each of above mentioned segmentation methods have been studied in detail in the literature. Fixed length page segmentation is simple and less complex in terms of implementation but the major problem with this approach is that it doesn't consider any semantics of the page while segmenting.

In DOM base page segmentation, the HTML tag tree's Document Object Model would be used while segmenting. An arbitrary passages based approach is given in [8]. Vision based page segmentation (VIPS) is in parallel lines with the way, humans views a page. VIPS [9] is a popular segmentation algorithm which segments a page based on various visual features.

Apart from the above mentioned segmentation methods a few novel approaches have been evolved during the last few years. An image processing based segmentation approach is illustrated in [10]. The segmentation process based text density of the contents is explained in [11]. The graph theory based approach to segmentation is presented in [12]. Repetition-based web page segmentation by detecting tag patterns for small-Screen Devices is explored in [13]. One of the approaches for web page segmentation for specific domains is detailed in [14]. A tree clustering based segmentation approach is provided in [15].

III. THE MODEL

This section elaborates about the mathematical model of the proposed system. The corresponding algorithm to carry out the task specified in the model is also explored in this section. The block diagram of the proposed model is as shown in Figure 1. It contains the following components: Page Segmentor:

This component is responsible for segmenting the contents of the page in to logically relevant units. Personalizer: This component handles the personalization of filtering. The Personalizer holds the profile-bag which contains user preferences. Segment Filter: Segment filter is another component in the model which handles individual segments and decides whether this segment should be incorporated in the filtered page or not.

3.1 Mathematical Model

In the proposed model each page that the user requests need to be segmented for filtration. Let us denote the source page byΦ .

The source page Φ has to be segmented in to various logically coherent parts. The source page Φ would be mapped as a DOM (Document Object Model) tree. The individual nodes of the DOM tree are processed by parsing the tree. The "block level" and "non-block level" nodes are identified and they are used as the building block of the individual segments.

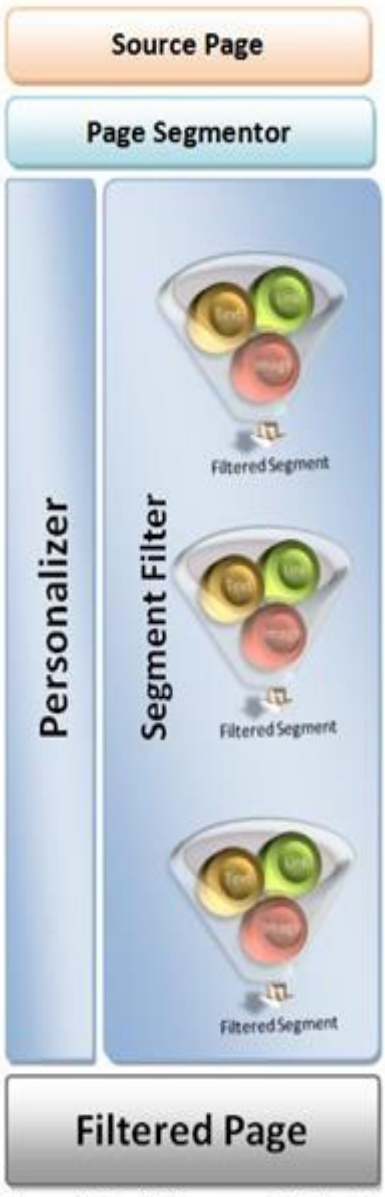

Figure 1. Block Diagram of the Model

The approach followed in this paper also incorporates the densitometry concepts in the segment building process. The densitometry considers the density of text present at a block unit in performing the segmentation process. As a result of the above mentioned process, the source page Φis segmented in to various units as shown in (1). Φ = { 1 , μ μ μ μ 2, 3... n } (1) The segmentation process shown in (1) is performed by the "Page Segmentor" component in the proposed model

IV. RESULTS AND DISCUSSIONS

η represent the text elements present in the segment under consideration; [ ] 1 2 , ,.. q κ κ κ represent the individual links presents in the segment and [ ] 1 2 , ... r λ λ λ represent the image elements present in the segments. The individual segments need to be processed for each of these three components to decide whether this segment can be allowed for display or it needs to be blocked.

In order to perform this, segment filter component includes three sub-components

a) Text Filter,

b) Link Filter and

c) Image Filter.

The focus of this research work is on the effect of segmentation and personalization. The actual filtration process can be either simple keyword based or it can be customized according to the requirements of implementation.

The proposed model incorporates personalization aspect. The user can configure the filter according to his/her requirements. The user preferences are represented using "Profile Bag". The profile bag involves two different tracks. These tracks are "Like Track" and "Un-Like Track". The block diagram of profile-bag is as shown in Figure 2. The figure consists of three horizontal layers.

The top layer denotes the overall profile-bag. The middle one represents the "Like-Track" and "Un-Like Track". The bottom layer in the Figure 2 denotes the keywords which form the "Like-Track" and "Un-Like Track". The profile bag is represented in the model asΓ.

The two different tracks of Γare represented as shown in (4). ω σ Γ= (4) In (4) ω represent the "Like Track" and σrepresent the "Un-Like Track" of the profile bag. Both ωand σcontains keywords that represent the user preferences.

The keywords in ωadds a positive booster and the keywords in σadds a negative booster. The filtration process can be represented as shown in (5). As a result of (5) the Text Weight, Link Weight and Image Weight are calculated as the sum of number of terms common between ωand elements giving a "+1" weight and number of terms common between σand elements giving a " 1" weight

Figure 2. The User Profile - Bag

If the sum of weights of all these three components exceeds a threshold level the segment is displayed otherwise it is blocked. ( ) : i i z if else μ δ μ μ □ □ ∀ ∈Φ Ψ+Λ+Θ ≥ ΦU □ □ Φ=□ □ ΦU □ □ □ □ (6) In (6), Φrepresents the filtered page in which segments whose weight has been calculated above the threshold limit are incorporated. When the weight is less than the threshold then a dummy segment z μ holding the message "segment blocked" would be added to the page. Figure 2.

The User Profile - Bag The dummy segment which would replace the filtered segment can be custom defined. The proposed model has another feature called "link hiding". In the case of link hiding, if the content to be blocked is having a hyperlink, instead of removing the content, the hyperlink alone can be removed which creates the similar impact as removing the content. 3.2 The Algorithm The algorithmic representation of the steps involved in the above explained model is explored in this section.

3.2 The Algorithm

The algorithmic representation of the steps involved in the above explained model is explored in this section.

Algorithm SegmentFilter

Input: Source Web Page Φ , profile bag Γ

Output : Filtered Page $\underline{\Phi}$

Begin

Segment the source page using page segmentor

$\Phi = \{\mu_1, \mu_2, \mu_3 ... \mu_n\}$

Initialize $\underline{\Phi}$ to NULL

For each segment μ_i

begin

Parse the segment μ_i into components $\{\Psi, \Lambda, \Theta\}$

Calculate Text weight $|\Psi| = \text{TF}(\Psi / \Gamma)$

Calculate Link Weight $|\Lambda| = \text{LF}(\Lambda / \Gamma)$

Calculate Image Weight $|\Theta| = \text{IF}(\Theta / \Gamma)$

If $(|\Psi|+|\Lambda|+|\Theta|) \geq \delta$ then

$\underline{\Phi} = \underline{\Phi} \cup \mu_i$

Else

$\underline{\Phi} = \underline{\Phi} \cup \mu_z$

End

Return ($\underline{\Phi}$)

End

V. EXPERIMENTS AND RESULT ANALYSIS

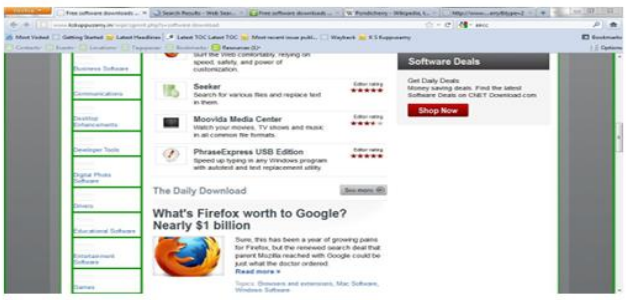

Figure 3. The Source Page

The page segments are filtered out based on the filtering preferences set up. The resultant page as shown in Figure 4.

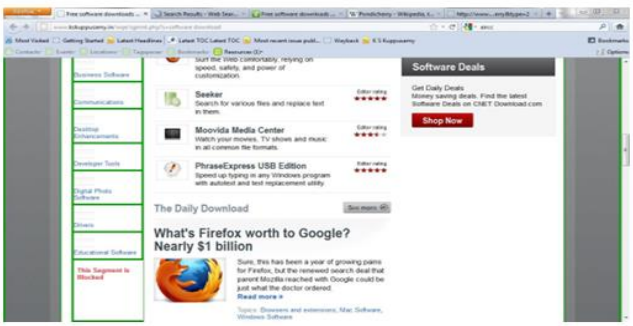

Figure 4. The Page after filtering the unwanted segments.

"Games" are filtered out as per the filtering preferences set. The contents of Table 1 list out the experimental results conducted on the proposed content filtering model. In the Table 1, MSC indicates the mean segment count, MFSC stands mean filtered segment count, MFP is mean false positives and MFN is mean false negative.

Session ID	MSC	MFSC	MFP	MFN	Accuracy (%)
1	27.52	5.2	1.2	1.5	90.189
2	30.25	3.5	0.8	1.2	93.388
3	43.53	4.5	1.3	1.3	94.027
4	20.67	2.7	0.7	0.2	95.646
5	18.45	1.5	1.6	0.4	89.16
6	14.66	2.3	3.5	0.5	72.715
7	16.78	4.3	3.1	1.1	74.97
8	17.67	1.3	1.2	1.5	84.72
9	14.85	1.8	0.5	0.8	91.246
10	25.52	2.6	0.9	0.9	92.947
11	12.45	5.2	0.9	1.3	82.329
12	22.15	5.1	0.6	1.4	90.971
13	23.45	3.9	1.1	0.6	92.751
14	25.45	4.2	1.2	0.8	92.141
15	12.45	4.3	1.8	1.1	76.707

The chart in Figure 5 compares the average number of segments filtered out in a session, the false positives and the false negatives. It can be observed that the mean of MFSC across the session is 3.49, whereas the mean of MFP and MFN are 1.3 and 0.9 respectively.

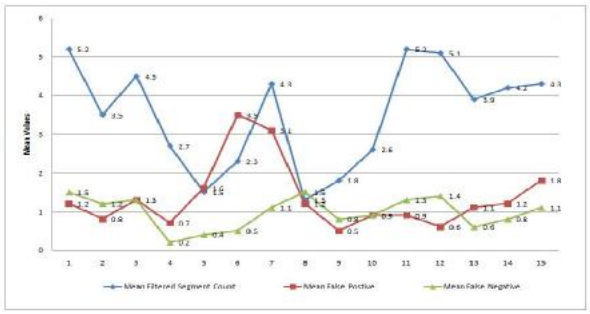

Figure 5. Comparison of MFSC, MFP and MFN

The proposed model has been implemented as prototype for experimentation. The prototype implementation is done with the software stack including Linux, Apache, MySql and PHP. For client side scripting JavaScript is used. With respect to the hardware, a Core i3 processor system with 3 GHz of speed, 8 GB of RAM is used. The internet connection used in the experimental setup is a 128 Mbps leased line. The screenshots of the prototype implementation are as shown in the Figure 3 and Figure 4. The screenshot shown in Figure 3 is of the original source page.

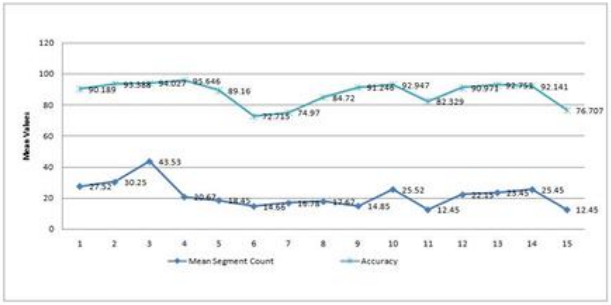

Figure 6. Comparison of MSC and Accuracy

VI. CONCLUSION

The proposed model for page filtering using segmentation and personalization renders the following advantages: • Instead of blocking the entire page in cases where the content to be blocked is present only at a portion of the page, the proposed model provides a distinct benefit to user. • Incorporation of personalization in the blocking process provides a tailor made content filtering system based on the user's needs. The future directions for this research work are as listed below: • In the proposed model the image filtering happens using the "alt" text provided with the image. In the future implementations some of the image analysis modules can be incorporated to make the image filtering much more efficient. • Incorporation of the capability to handle languages other than English would make the system more efficient in the cases of non-English web pages.

REFERENCES

[1]. P. Nirmala, T. Manimegalai, J. R. Arunkumar, S. Vimala, G. Vinoth Rajkumar, Raja Raju, "A Mechanism for Detecting the Intruder in the Network through a Stacking Dilated CNN Model", Wireless Communications and Mobile Computing, vol. 2022, Article ID 1955009, 13 pages, 2022. https://doi.org/10.1155/2022/1955009.

[2]. D. Sathyanarayanan, T. S. Reddy, A. Sathish, P. Geetha, J. R. Arunkumar and S. P. K. Deepak, "American Sign Language Recognition System for Numerical and Alphabets," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-6, doi: 10.1109/RMKMATE59243.2023.10369455.

[3]. J. R. Arunkumar, Tagele berihun Mengist, 2020" Developing Ethiopian Yirgacheffe Coffee Grading Model using a Deep Learning Classifier" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4, February 2020. DOI: 10.35940/ijitee.D1823.029420.

[4]. Ashwini, S., Arunkumar, J.R., Prabu, R.T. et al. Diagnosis and multi-classification of lung diseases in CXR images using optimized deep convolutional neural network. Soft Comput (2023). https://doi.org/10.1007/s00500-023-09480-3

[5]. J.R.Arunkumar, Dr.E.Muthukumar," A Novel Method to Improve AODV Protocol for WSN" in Journal of Engineering Sciences" ISSN NO: 0377-9254Volume 3, Issue 1, Jul 2012.

[6]. R. K, A. Shameem, P. Biswas, B. T. Geetha, J. R. Arunkumar and P. K. Lakineni, "Supply Chain Management Using Blockchain: Opportunities, Challenges, and Future Directions," 2023 Second International Conference on Informatics (ICI), Noida, India, 2023, pp. 1-6, doi: 10.1109/ICI60088.2023.10421633.

[7]. Arunkumar, J. R. "Study Analysis of Cloud Security Chanllenges and Issues in Cloud Computing Technologies." Journal of Science, Computing and Engineering Research 6.8 (2023): 06-10.

[8]. J. R. Arunkumar, R. Raman, S. Sivakumar and R. Pavithra, "Wearable Devices for Patient Monitoring System using IoT," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 381-385, doi: 10.1109/ICCES57224.2023.10192741.

[9]. S. Sugumaran, C. Geetha, S. S, P. C. Bharath Kumar, T. D. Subha and J. R. Arunkumar, "Energy Efficient Routing Algorithm with Mobile Sink Assistance in Wireless Sensor Networks," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10201142.

[10]. R. S. Vignesh, V. Chinnammal, Gururaj.D, A. K. Kumar, K. V. Karthikeyan and J. R. Arunkumar, "Secured Data Access and Control Abilities Management over Cloud Environment using Novel Cryptographic Principles," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10199616.

[11]. Syamala, M., Anusuya, R., Sonkar, S.K. et al. Big data analytics for dynamic network slicing in 5G and beyond with dynamic user preferences. Opt Quant Electron 56, 61 (2024). https://doi.org/10.1007/s11082-023-05663-2

[12]. Krishna Veni, S. R., and R. Anusuya. "Design and Study Analysis Automated Recognition system of Fake Currency Notes." Journal of Science, Computing and Engineering Research 6.6 (2023): 16-20.

[13]. V. RamKumar, S. Shanthi, K. S. Kumar, S. Kanageswari, S. Mahalakshmi and R. Anusuya, "Internet of Things Assisted Remote Health and Safety Monitoring Scheme Using Intelligent Sensors," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10199766.

[14]. R. S. Vignesh, R. Sankar, A. Balaji, K. S. Kumar, V. Sharmila Bhargavi and R. Anusuya, "IoT Assisted Drunk and Drive People Identification to Avoid Accidents and Ensure Road Safety Measures," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10200809.

[15]. I. Chandra, G. Sowmiya, G. Charulatha, S. D, S. Gomathi and R. Anusuya, "An efficient Intelligent Systems for Low-Power Consumption Zigbee-Based Wearable Device for Voice Data Transmission," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10083856.

[16]. G. Karthikeyan, D. T. G, R. Anusuya, K. K. G, J. T and R. T. Prabu, "Real-Time Sidewalk Crack Identification and Classification based on Convolutional Neural Network using Thermal Images," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 1266-1274, doi: 10.1109/ICACRS55517.2022.10029202.

[17]. R. Meena, T. Kavitha, A. K. S, D. M. Mathew, R. Anusuya and G. Karthik, "Extracting Behavioral Characteristics of College Students Using Data Mining on Big Data," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10084276.

[18]. S. Bharathi, A. Balaji, D. Irene. J, C. Kalaivanan and R. Anusuya, "An Efficient Liver Disease Prediction based on Deep Convolutional Neural Network using Biopsy Images," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1141-1147, doi: 10.1109/ICOSEC54921.2022.9951870.

[19]. I. Chandra, G. Sowmiya, G. Charulatha, S. D, S. Gomathi and R. Anusuya, "An efficient Intelligent Systems for Low-Power Consumption Zigbee-Based Wearable Device for Voice Data Transmission," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10083856.

[20]. Revathi, S., et al. "Developing an Infant Monitoring System using IoT (INMOS)." International Scientific Journal of Contemporary Research in Engineering Science and Management 6.1 (2021): 111-115.

[21]. J.R.Arunkumar, Dr.E.Muthukumar,‖ A Novel Method to Improve AODV Protocol for WSN‖ in Journal of Engineering Sciences‖ ISSN NO: 0377-9254Volume 3, Issue 1, Jul 2012.

[22]. R. S. Vignesh, A. Kumar S, T. M. Amirthalakshmi, P. Delphy, J. R. Arunkumar and S. Kamatchi, "An Efficient and Intelligent Systems for Internet of Things Based Health Observance System for Covid 19 Patients," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ICECONF57129.2023.10084066.

[23]. I. Chandra, K. V. Karthikeyan, R. V, S. K, M. Tamilselvi and J. R. Arunkumar, "A Robust and Efficient Computational Offloading and Task Scheduling Model in Mobile Cloud Computing," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ICECONF57129.2023.10084293.

[24]. R. K, A. Shameem, P. Biswas, B. T. Geetha, J. R. Arunkumar and P. K. Lakineni, "Supply Chain Management Using Blockchain: Opportunities, Challenges, and Future Directions," 2023 Second International Conference on Informatics (ICI), Noida, India, 2023, pp. 1-6, doi: 10.1109/ICI60088.2023.10421633.

[25]. J. R. Arunkumar, and R. Anusuya, "OCHRE: A Methodology for the Deployment of Sensor Networks." American Journal of Computing Research Repository, vol. 3, no. 1 (2015): 5-8.

JSCER