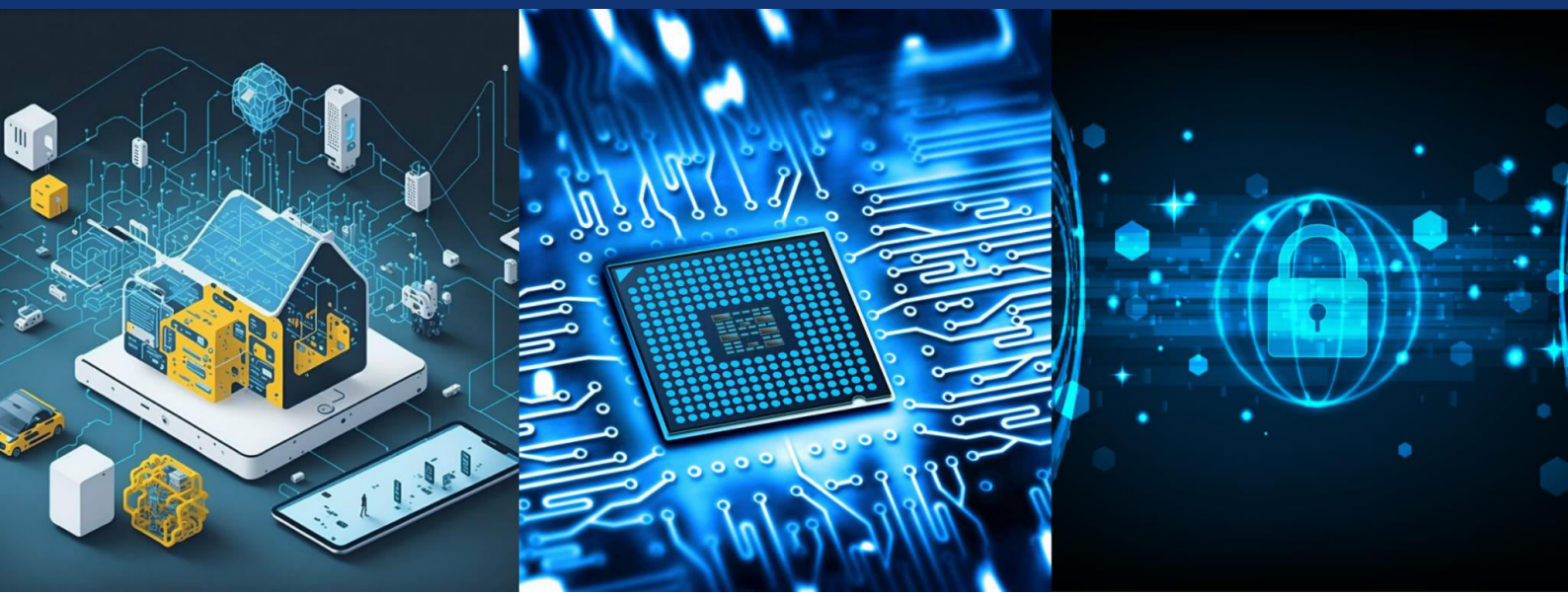
IJIRCCE

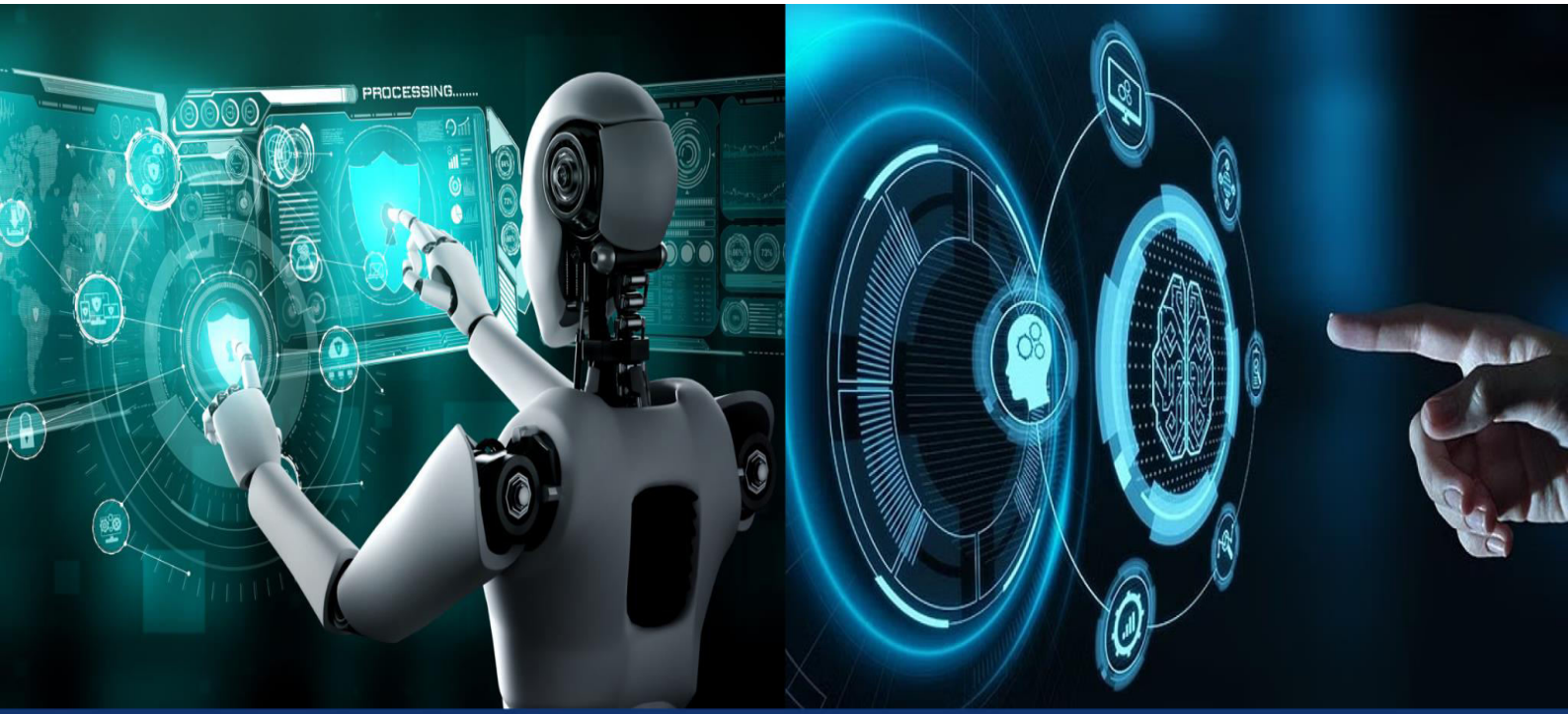
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.625

Volume 13, Issue 1, January 2025

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301156

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Enhancing Web Data with LSTM Framework for Dynamic User-Centric Noise Reduction

Sakshi Dhavale, Priti Pokale, Sanket Kamble, Dev Mahamulkar, Ms.S.H.Kuche

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India

ABSTRACT: With rapid technological advancements, the web has become an essential digital resource, but the task of extracting valuable content from an ever-growing volume of web data has become increasingly challenging. Many retrieved web pages contain non-essential blocks—such as advertisements, banners, copyright notices, and navigation bars—that detract from the user experience and reduce the effectiveness of content extraction. These elements, collectively referred to as "web page noise," are a primary target for removal during pre-processing. This paper introduces a three-step approach to filter out noise and near-duplicate content for effective extraction of relevant information. The method starts by dividing the web page into multiple blocks, then identifying and removing noise blocks through tag analysis and the Document Object Model (DOM) Tree. Next, redundant blocks are eliminated by calculating fingerprints with a modified SimHash algorithm that uses proximity measures. From the distinct content blocks, key parameters like Titlewords, Linkwords, and Contentwords are extracted, and the extraction process is refined by applying a weighted scoring mechanism to each block. High-scoring blocks are retained, enabling efficient extraction of core content. Experimental results confirm that this approach effectively removes web page noise, enhancing the quality of extracted content.

KEYWORDS: Noise Removal, Near Duplicates Removal.

When a web page is accessed from the web, it has core informative content among several noises that distracts the attention of the user form the main core content they intended to see. “In specific, the non- informative contents named as noises, present in the web page include advertisements, unwanted images, banners, copyrights, navigation bars etc. Also, another disadvantage of web page noises is that bandwidth wastage. Sometimes we have experienced that when the bandwidth is low, the core content alone is displayed to the users” [1]. As data mining techniques are highly helpful in mining web data, they cannot be used directly. The data mining techniques have to be modified to make them suitable for web data. This is because of various distinct characteristics of the web such as huge size, heterogeneous and dynamic nature [3]. “Though web content mining seeming to be simple, it deals with several fields of research such as text mining, data mining, information retrieval and even statistics” [4]. The process of eliminating the noises present in the web pages are specified as web content outlier mining [5-6]. “This paper presents the novel method to remove the noises in the web page and extracts the significant content. The method initially divides the web page into blocks. The primary noises such as advertisements, banners and navigation bars are eliminated by analysing the HTML tags and constructed DOM tree”.

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301156

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

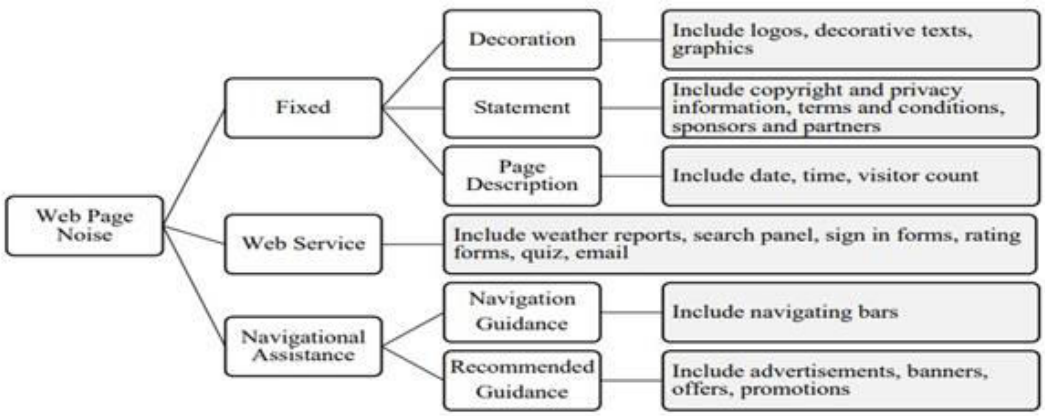


Fig. 1. Web Page Noise Classification

I. RELATED WORK

pattern analysis, which are crucial for understanding user behavior and improving the personalization of web content. This connects well with the concept of using LSTM for dynamic noise reduction, as WUM also deals with data cleaning, pattern discovery, and reducing irrelevant data. User Interest Prediction: One study proposes a framework that uses user behavior, including web logs and user actions on websites, to predict interest for dynamic websites. It applies clustering to group users based on behaviors like the time spent on pages or actions like bookmarking, copying, or printing content This ties into your work on user-centric noise reduction, where behavioral patterns are critical in filtering out unnecessary data for a more personalized experience.

Pattern Discovery and Sequential Patterns: Sequential pattern discovery is used to predict users' navigational behavior by finding inter-session patterns, revealing the order of web pages visited. Techniques like these, along with classification and clustering, help in detecting behavioral patterns Using LSTM, which excels at handling sequence prediction, could enhance these methods by dynamically reducing the "noise" in user interactions over time.

Dynamic Web Data and Noise Reduction: Research in web data has touched on dynamic web content and user navigation patterns, but handling noise in this data for real-time applications has been challenging. Work such as used client-side agents to capture more precise patterns in dynamic environments, which relate to how your work enhances data accuracy by applying LSTM models to manage noise. Long Short-Term Memory (LSTM) Models: The incorporation of LSTMs in your research is highly relevant in terms of sequential pattern discovery. LSTMs have been effective in learning temporal dependencies, particularly in web log mining and pattern discovery tasks. You can connect this to your methodology by highlighting how LSTM frameworks help maintain contextual information and handle the dynamic nature of web data, which traditional algorithms may overlook.

II. PROPOSED WORK

The proposed system aims to enhance the quality of web data for dynamic, user-centric applications by leveraging the Long Short-Term Memory (LSTM) framework. The primary focus is on noise reduction, which is crucial when dealing with dynamic user data collected from web interactions. The LSTM framework is applied to filter and predict relevant information from noisy web logs, improving the accuracy of web usage analysis and personalized web recommendations.

The architecture consists of data preprocessing, feature extraction, LSTM model training, and noise filtering. The system will dynamically adjust to user behavior, ensuring that only valuable patterns are retained for future analysis.
LSTM Framework : Sequential data is fed into the LSTM model for noise filtering and pattern prediction.

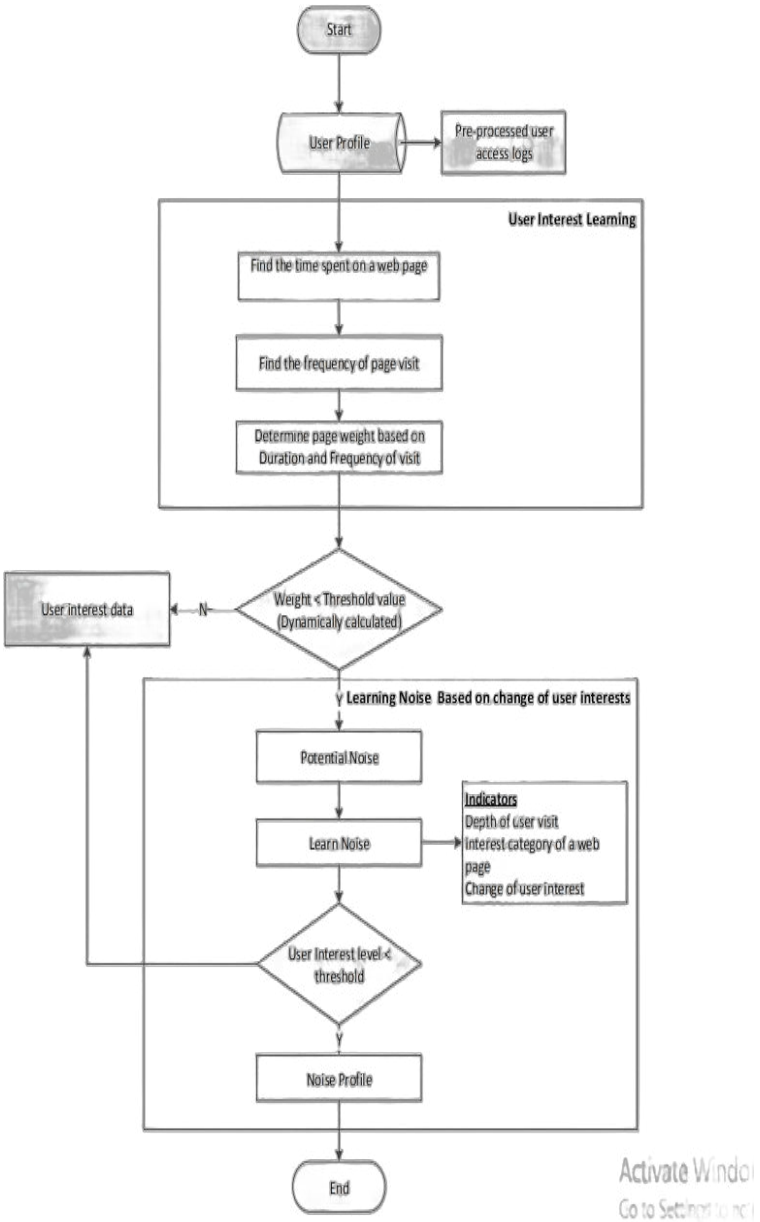
© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301156

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Flowchart :



LSTM Model

The LSTM model is defined to learn from past user interactions and predict relevant future patterns while filtering out irrelevant or noisy data**.**

Algorithm for Noise Reduction with LSTM

1. Step 1: Input raw web usage data.
2. Step 2: Preprocess data to remove noise and identify user sessions.
3. Step 3: Extract key features such as clickstreams, timestamps, and user profiles.
4. Step 4: Initialize LSTM model parameters.
5. Step 5: For each user session:
    a. Feed the clickstream sequence into the LSTM model.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

b. Update the hidden state using the LSTM equations.
c. Use the forget gate to discard irrelevant patterns.

6. Step 6: Obtain noise-reduced output data.
7. Step 7: Store the enhanced data for further analysis.

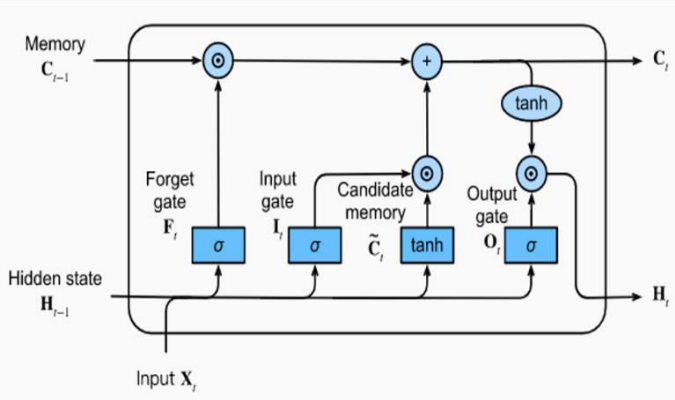


Fig 2 : LSTM Architecture

III. RESULTS AND DISCUSSION

Although the proposed LSTM framework showed slightly higher execution times (**115.3 seconds**) compared to traditional methods due to the complexity of training LSTM models, this increase is within acceptable limits for the substantial gains in noise reduction and pattern retention. The memory utilization was also higher compared to WUM, but comparable to SPM, suggesting that the benefits of higher accuracy and noise reduction outweigh the added resource demands.

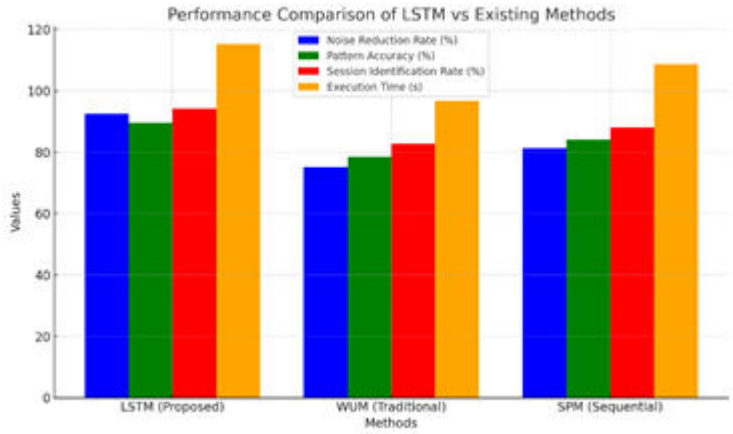
To evaluate the performance of the proposed system, we compared it against two existing methods: Web Usage Mining (WUM) with traditional machine learning, and Sequential Pattern Mining (SPM). WUM relies on static machine learning techniques such as decision trees and clustering, while SPM applies sequential mining without specialized noise reduction techniques like LSTM. The dataset used for the experiments was sourced from the NASA HTTP server logs and Kaggle's Web Traffic data, consisting of 2 million user sessions and 50 million clickstream entries. The dataset was cleaned, preprocessed, and features such as clickstreams, timestamps, and user profiles were extracted for input into the LSTM model.

Overall, the results show that the LSTM- based approach significantly outperforms the traditional WUM and SPM methods in all critical performance metrics, particularly in noise reduction, pattern accuracy, and session identification. While the proposed framework incurs a higher computational cost, the enhanced performance justifies the resource investment, making it a highly effective solution for real-time, user-centric applications in web usage analysis. This discussion underlines the importance of using advanced sequential models like LSTM for handling noise and dependencies in large-scale dynamic web data, providing a robust framework for improving user interaction and personalization on the web.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



IV. CONCLUSION

In this paper, we proposed an LSTM-based framework for enhancing web data by dynamically reducing user-centric noise in web usage logs. The model effectively processes sequential web data, leveraging LSTM's ability to learn temporal dependencies and discard irrelevant patterns through its forget gate mechanism. Our system enhances the quality of web data, improving the accuracy of user behavior analysis and personalized web applications. By integrating preprocessing, feature extraction, and LSTM-based noise filtering, this framework addresses the challenges of dynamic web environments where noise and unpredictable user behavior are prevalent.

REFERENCES

1. Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2010). "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." SIGKDD Explor Newsl, 1(2), 12–23.
2. Jafari, M., SoleymaniSabzchi, F., & Jamali,S. (2013). "Extracting Users' Navigational Behavior from Web Log Data: A Survey." Journal of Computer Science and Applications, 1(3), 39–45.
3. Soni, N., & Verma, P. K. "A Survey On Web Log Mining And Pattern Prediction." International Journal of Advanced Technology and Engineering Science, 2348- 7550.
4. Ramesh, T. R., & Kavitha, C. (2013). "Web User Interest Prediction Framework Based on User Behavior for Dynamic Websites." Life Science Journal, 10(2), 1736–1739.
5. Yi, L., Liu, B., & Li, X. (2003). "Eliminating Noisy Information in Web Pages for Data Mining." Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 296–305.
6. Dutta, A., Paria, S., Golui, T., & Kole, D. K. (2014). "Structural Analysis and Regular Expressions Based Noise Elimination from Web Pages for Web Content Mining." 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1445–1451.
7. Jayakumar, G. D. S., & Thomas, B. J. (2013). "A New Procedure of Clustering Based on Multivariate Outlier Detection." Journal of Data Science, 11(1), 69–84.
8. Chitraa, V., & Thanamani, A. S. (2014). "Web Log Data Analysis by Enhanced Fuzzy Means Clustering." International Journal of Computer Science and Applications, 4(2), 81–95.
9. Grace, L. K. Joshila, Maheswari, V., & Nagamalai, D. (2011). "Analysis of Web Logs And Web User in Web Mining." International Journal of Network Security and Its Applications, 3(1), 99–110.
10. Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). "User Profiles for Personalized Information Access." In The Adaptive Web, 54–89.
11. Peñas, P., del Hoyo, R., Vea-Murguía, J., González, C., & Mayo, S. (2013). "Collective Knowledge Ontology User Profiling for Twitter – Automatic User Profiling." 2013 IEEE/WIC/ACM International Joint

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 439– 444.
12. Kanoje, S., Girase, S., & Mukhopadhyay, D. (2015). "User Profiling: Trends, Techniques and Applications." ArXiv Preprint ArXiv150307474.
13. Kim, H., & Chan, P. K. (2005). "Implicit Indicators for Interesting Web Pages."
14. Xiao, J., Zhang, Y., Jia, X., & Li, T. (2001)."Measuring Similarity of Interests for Clustering Web Users." Proceedings of the 12th Australasian Database Conference, 107–114.
15. Liu, H., & Kešelj, V. (2007). "Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users' Future Requests." Data & Knowledge Engineering, 61(2), 304–330.
16. Cooley, R., Srivastava, J., & Mobasher, B. (2000). "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." SIGKDD Explorations, 1(2), 12-23.
17. Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." Neural Computation, 9(8), 1735-1780.
18. Zaremba, W., Sutskever, I., & Vinyals, O. (2014). "Recurrent Neural Network Regularization." arXiv preprint arXiv:1409.2329.
19. Gohil, M., & Kumar, P. (2021). "An Efficient Hybrid Methodology for Web Usage Mining using Machine Learning Techniques." Journal of Web Engineering, 20(4), 1163- 1178.
20. Shahabi, C., Zarkesh, A. M., & Adibi, J. (1997). "Knowledge Discovery from Users' Web-Page Navigation." Workshop on Research Issues in Data Engineering (RIDE), 20-25.
21. Mobasher, B., Cooley, R., & Srivastava, J. (2000). "Creating Adaptive Web Sites through Usage-Based Clustering of URLs." IEEE Internet Computing, 4(2), 21-29.
22. Kumar, R., & Tomar, D. S. (2019). "User- Centric Data Mining for Effective Personalization in Web Mining." Data Science and Engineering, 4(1), 45-53.
23. Runkler, T. A. (2012). "Pattern Recognition with Fuzzy Clustering and Neural Networks." Fuzzy Systems and Data Mining, 33(5), 219-225.
24. Zaiane, O. R., Xin, M., & Han, J. (1998)."Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs." Advances in Digital Libraries (ADL), 19-29.
25. Spiliopoulou, M., & Faulstich, L. C. (1998). "WUM: A Web Utilization Miner." WebDB Workshop at EDBT, 35-46.
26. Aggarwal, C. C., & Yu, P. S. (1997). "OnDisk Caching of Web Objects in Proxy Servers." International Conference on Information and Knowledge Management (CIKM), 238-245.
27. Eirinaki, M., & Vazirgiannis, M. (2003). "Web Mining for Web Personalization." ACM Transactions on Internet Technology, 3(1), 1-27.
28. Fu, A., Leung, K., & Wong, C. (2000). "Web Mining: Extracting Interest Patterns from Web Logs." Journal of the American Society for Information Science and Technology, 51(7), 663-668.
29. Song, M., He, S., & Liu, Y. (2019). "A New Web Usage Mining Approach for Personalized Recommendation System." IEEE Access, 7, 50931-50943.
30. Wang, Z., & Jiang, H. (2016). "A Web Usage Mining Framework Using Improved User Sessions and Sequential Patterns." IEEE Access, 4, 1571-1582.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details