INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 10, Issue 1 - V10I1-1286)
Available online at: https://www.ijariit.com

Noise reduction in web data – A learning approach based on dynamic user interest

Sakshi Balbansi
sakshibalbansi@gmail.com
JSPM's Bhivrabai Sawant Institute of Technology and Research, Wagholi-Pune.

Aditya Pathak
pathakaditya.cu@gmail.com
JSPM's Bhivrabai Sawant Institute of Technology and Research, Wagholi-Pune.

Akanksha Memane
memaneak18@gmail.com
JSPM's Bhivrabai Sawant Institute of Technology and Research, Wagholi-Pune.

Mrunmayi Khaamkar
mrunmayikhaamkar1@gmail.com
JSPM's Bhivrabai Sawant Institute of Technology and Research, Wagholi-Pune.

Prof. Vijay Sonawane
sonawanevijay4@gmail.com
JSPM's Bhivrabai Sawant Institute of Technology and Research, Wagholi-Pune.

ABSTRACT

An advanced noise reduction technique harnessing the power of Long Short-Term Memory (LSTM) networks is introduced to tackle the issue of noise in web data. In contrast to traditional methods employed for web data noise reduction, which may grapple with complexities such as large network depths and training inefficiencies, this novel approach takes a fresh perspective. Initially designed for DE noising natural datasets, this LSTM-based algorithm has been carefully adapted and fine-tuned to specifically target noise reduction in web data. Through meticulous parameter adjustments and extensive experimentation, we have successfully demonstrated the effectiveness of LSTM in removing noise from web data, achieving high levels of efficiency.Our thorough analysis and comparative experiments underscore the potential and viability of the LSTM-based approach in the domain of web data noise reduction. This algorithm not only holds promise but also signifies its importance in advancing the field of web data processing and analysis, marking a significant step forward in enhancing data quality for web-related applications.

Keywords: *Machine Learning, Noise Web data learning data, User Interest, Web user profile, Web log data.*

I. INTRODUCTION

In today's digital environment, the ubiquity of the Internet has made it an essential part of our daily lives, and most users search for important information [1]-[3]. However, the large amount of irrelevant information on the Internet [4] poses a significant challenge in ensuring content is relevant to specific users. This irrelevant web content is often referred to as “noise” and contains information that does not constitute the main content of the web page [5] , [6] . Network data denoising involves detecting and removing network data that is irrelevant to the content of the home page or not useful to a particular user [7].

Although existing research confirms that the reduction in web information is site-specific and involves the removal of external web pages unrelated to the main content, it is primarily focused on analyzing data outside the web to interpret data on the web. important content. Web User Profile [8]. In this case, noise is not limited to ads from external pages, links, dead URLs or devices that do not constitute the main content of the page. It also contains important information that does not reflect changes in customer preferences.
This process, which uses various machine learning tools and algorithms to extract important data from network data, is called network or data mining [1], [2]. This technique is to analyze user interest from the web log file, which contains information about the user's relevant website [9]. Website logs may provide information about user interests, including basic information such as IP addresses, user visit times, searches, pages visited on the web, and time spent on each web page.

The terms "blog files" and "Web data" are used interchangeably in this study because log files mainly contain Web data. Therefore, the removal of popular data in the network is based on user dataset data extraction analysis. However, it is worth noting that in practice it is very difficult to create a user network profile without popular information. An Internet user profile is a description of the interests, characteristics, and preferences of users of a particular website [10]-[12]. User interests can be explicit, in that users communicate their interests and opinions about information on the web, or implicit, in that the system infers those users' interests through various means, such as the frequency and time of visiting the web page. [14], [15]]. Many users will be reluctant to share their true needs with information on the website, so targeting dissatisfied users is an important consideration.

Previous research on data network noise reduction was based on the assumption that the data network is static [16]. For example, [17] and [18] proposed a method to match noise patterns detected from web pages to store noise data for classification and later removal. This approach means that removing noise from data networks is based on previous noise data. However, in the context of data transfer over the web, this model will be incomplete, emphasizing the importance of considering customer satisfaction pressure [19], [20]. Web access patterns are dynamic not only due to constant changes in Web content, but also due to changes in consumer preferences [21]. For example, Internet users celebrate weddings, Christmas, birthdays, etc. may show more interest in information about events. Therefore, it is necessary to investigate how the changes affect the network data removal process.

To solve the dynamic problem of reducing network data noise, this work firstly presents a machine learning algorithm that can learn and identify noise in network data. The proposed algorithm takes into account the change in user interest and the development of the knowledge network to know and learn popular knowledge well. The main innovations of this study are: Highlighting the impact of customer satisfaction and network data updating on the noise removal process, including contributions and limitations of the current research.

Introduce a machine learning algorithm that learns from user noise to remove historical data, including changes in customer preferences and changing network data.
Practical use of the proposed tool should reduce the noise in removing important information from the user's web profile, which can improve the quality of user experience.

II. LITERATURE REVIEW

In first study, "Web Data Denoising Using LSTM-Based Autoencoders" this paper introduces an LSTM-based autoencoder architecture for web data denoising. The model is trained on a large web dataset and demonstrates significant improvements in noise reduction. This model is made of keywords such as artificial neural networks,LSTM cells, Integrated squared error, internal model control.Main motive of this project is to control any industrial process till the data from the process to be controlled are available.This project also focuses in reducing the complexity in control structure design process.[22]

In secondly "Efficient Noise Reduction in Web Data with LSTM Networks". This research explores the efficiency of LSTM networks in reducing noise in web data. It proposes an optimized architecture for faster and more effective denoising.The architecture is made of Data collection,denoising,evalution,pre-processing, skull stripping, Bi-LSTM. It is main motive to reduce noise from image and maintain the quality of image.[23]

In the third study paper "Adaptive Noise Reduction for Web Text Data using LSTM" .The paper presents an adaptive noise reduction approach using LSTM networks specifically designed for web text data. It adapts to varying levels of noise and achieves improved accuracy in text classification tasks. Architecture are build by methods such as data processing, attention with encoder and decoder, attention mechanism model, model evaluation matrices, optimizer selection and parameter optimization.[24]

In this paper study "Web Data Preprocessing: A Comparative Study of LSTM and CNN".
This study compares LSTM and CNN-based methods to remove data network noise. It evaluates their performance in processing different types of mixed audio data and discusses their relative effectiveness. Algorithms used include deep neural network (DNN), convolutional neural network (CNN), recursive neural network (RNN), LSTM algorithm and hypothesis analysis. These models often focus on emotional analysis using deep learning. [25]

Also this article "Semi-View Noise Reduction in Web Images with LSTM". For online image files, this article shows a method to reduce noise by half. It uses labeled and unlabeled data to improve denoising. To reduce image noise, a neural network is used to perform image denoising using semi-supervised learning. Semi-supervised first draws the tensor. According to qualitative and quantitative observations, the algorithm can store the inequality of the tensor, then identify the prior distribution of the block group according to the Gaussian mixture, and finally identify it using Bayesian inference. final result. conclusion. Block the group. [26]

III. PROBLEM STATEMENT

Noise reduction web data using LSTM(Long-short-Term Memory)

The central challenge at hand is the development and implementation of a noise reduction system tailored for web data, and the chosen methodology involves the utilization of LSTM (Long Short-Term Memory) algorithms. Web data is inherently fraught with noise, stemming from diverse sources of inconsistencies, inaccuracies, and the presence of irrelevant information. This inherent noise significantly undermines the quality and reliability of data-driven applications and analytical processes that rely on web data sources. The overarching objective of this project is to engineer an automated solution capable of discerning and effectively filtering out this noise from web data. By doing so, the project aspires to elevate the usability and trustworthiness of web data, making it more amenable for deployment in downstream applications and analyses. In essence, the project endeavors to provide a robust and reliable foundation for the utilization of web data in data-driven decision-making, research, and various other domains by mitigating the detrimental impact of noise on the data's integrity and utility.

IV. PROPOSED MEDTHODOLOGY

Project Modules: To provide a clear and concise overview of our proposed approach, we encapsulate the core concepts and research objectives in a visual representation, as illustrated in Figure 1. This visualization serves as a succinct summary of the foundational principles guiding this research initiative.

The project is structured into several key modules, each playing a vital role in enhancing the use and understanding of web data:
Web Data Extraction: This initial module lays the foundation by focusing on gathering web data from various sources and channels. This process may encompass techniques like web scraping, utilizing APIs, and employing other data retrieval methods to compile a comprehensive dataset.

Web User Profile Creation: Once the data is collected, the project proceeds to create web user profiles. These profiles are designed to encapsulate a user's preferences, behaviors, and interactions with web content, essentially providing a detailed snapshot of their online presence.

User Interest Learning: This module takes a machine learning approach to understand and learn user interests. By analyzing user interactions with web content, it identifies patterns and preferences, effectively deciphering what captures a user's attention and engagement.

Interest Level Determination on Visited Pages: This stage focuses on gauging the depth of user interest in web pages they visit. Various factors such as visit frequency, duration, recency, and the extent of exploration into links on the pages are evaluated to determine the level of engagement.

Web Data Classification using LSTM: Leveraging Long Short-Term Memory (LSTM) technology, this module creates a classification model. Its purpose is to categorize web data based on user interests. This segmentation effectively separates valuable content from noise, making it easier to serve relevant information to users.

Noise Learning in Web Data: Here, the project zeroes in on the identification and categorization of noise within web data. The LSTM model is instrumental in this process, learning to differentiate between content that adds value and irrelevant or disruptive noise.

User Profile Update: As users' interests and web data evolve over time, this module ensures that the user profiles remain current and reflective of their changing preferences and behaviors. It adapts continuously to maintain the relevance and accuracy of the profiles.
This comprehensive approach to web data analysis not only improves the quality of user experiences but also has potential applications in content recommendation systems, personalization, and enhancing the effectiveness of data-driven decision-making processes. It combines data collection, machine learning, and user profile management to provide a holistic solution for handling web data.

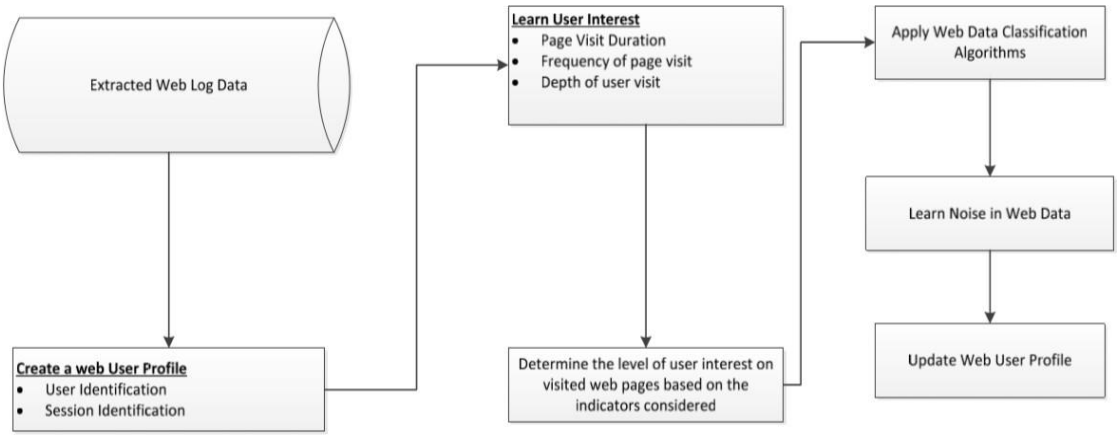

Figure No.1 Proposed System

LSTM ALGORITHMS

Applying short-term memory (LSTM) networks to denoise Web data is a good way to solve the problems caused by popular Web data. LSTM is a type of recurrent neural network (RNN) designed to solve the problem inherent in RNNs (the vanishing problem). The unique feature of LSTM is its ability to perfectly retain memory for extended arrays, making it particularly suitable for noise reduction in data networks.

The main advantage of LSTM is that it is not sensitive to variable length, which distinguishes it from other learning methods, including traditional RNN and hidden Markov models. He has a unique "short-term memory" that allows him to store important information thousands of times; This is especially important for removing noise in a data network.

The application of LSTM in reducing network data noise is diverse and effective. It can be used to classify, process and forecast data based on time series. The basic LSTM unit consists of a unit, an input gate, an output gate and a memory gate. These components control the flow of data in the network: The memory gate decides what data from previous states to keep or discard, making it easier to retain relevant background information while eliminating noise.

Data entry to determine what new data is relevant and should be retained in its current state helps maintain the accuracy and quality of data in the database. The output gate controls the data to be output, allowing the LSTM network to select relevant data, effectively reducing noise and maintaining long-term dependence on accurate prediction.

The main element of LSTM architecture is memory, which makes it different from traditional RNN. This storage unit has the unique ability to capture and store data in multiple arrays; This makes it particularly suitable for working with long-term addictions. In the field of data networks, LSTM's ability to separate important data from noise is important. LSTM improves network data quality by adjusting the data flow and selectively preserving relevant content while filtering out noise. This makes it useful in reducing data noise and ultimately improving the reliability and efficiency of data networks for various applications.

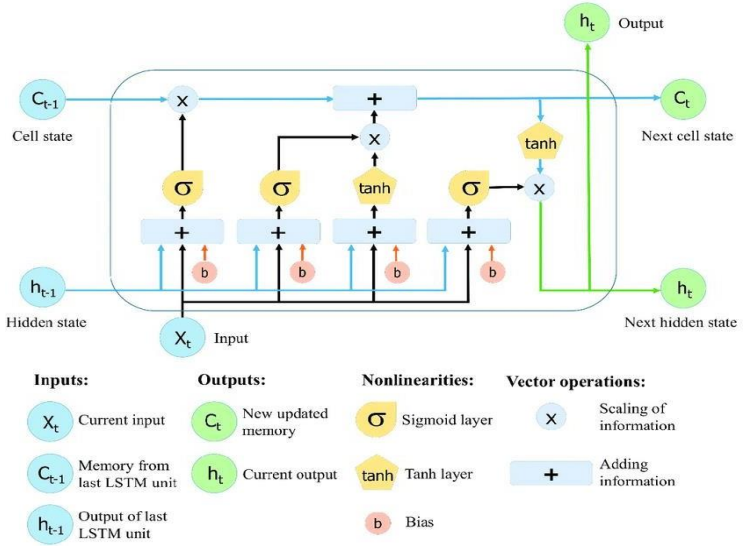

FIGURE NO. 2 ARCHITECTURE OF LSTM

V. CONCLUSION

In the era of the internet's information abundance, the ability to sift through the vast sea of web data and deliver meaningful, personalized content to users has become a critical necessity. Our proposed web data management system, comprising a series of interconnected modules, presents a comprehensive solution to this challenge. It begins with the initial extraction of web data and extends all the way to the continuous updates of user profiles. This approach ensures that users not only receive content that is relevant to their interests but also content that adapts to their evolving preferences. What sets our system apart is the incorporation of LSTM technology, which brings a layer of intelligence to the process, making it all the more effective.

VI. REFRENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data', SIGKDD Explor Newsl, vol. 1, no. 2, pp. 12–23, Jan. 2010.
[2] M. Jafari, F. SoleymaniSabzchi, and S. Jamali, 'Extracting Users'Navigational Behavior from Web Log Data: a Survey', J. Comput. Sci. Appl. J. Comput. Sci. Appl., vol. 1, no. 3, pp. 39–45, Jan. 2013.
[3] N. Soni and P. K. Verma, 'A Survey On Web Log Mining And Pattern Prediction', Int. J. Adv.Technol. Eng. Sci.-2348-7550.
[4] T. R. Ramesh and C. Kavitha, 'Web user interest prediction framework based on user behavior for dynamic websites', Life Sci. J., vol. 10, no. 2, pp. 1736–1739, 2013.

[5] L. Yi, B. Liu, and X. Li, 'Eliminating Noisy Information in Web Pages for Data Mining', in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 296–305.
[6] A. Dutta, S. Paria, T. Golui, and D. K. Kole, 'Structural analysis and regular expressions based noise elimination from web pages for web content mining', in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1445–1451.
[7] G. D. S. Jayakumar and B. J. Thomas, 'A new procedure of clustering based on multivariate outlier detection', J. Data Sci., vol. 11, no. 1, pp. 69–84, 2013.
[8] V. Chitraa and A. S. Thanamani, 'Web Log Data Analysis by Enhanced Fuzzy C Means Clustering', Int. J. Comput. Sci. Appl., vol. 4, no. 2, pp. 81–95, Apr. 2014
[9] L. K. Joshila Grace, V. Maheswari, and D. Nagamalai, 'Analysis of Web Logs And Web User In Web Mining', Int. J. Netw. Secur. Its Appl., vol. 3, no. 1, pp. 99–110, Jan. 2011
[10] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, 'User profiles for personalized information access', in The adaptive web, Springer, 2007, pp. 54–89.
[11] P. Peñas, R. del Hoyo, J. Vea-Murguía, C. González, and S. Mayo, 'Collective Knowledge Ontology User Profiling for Twitter – Automatic User Profiling', in 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, vol. 1, pp. 439–444.
[12] S. Kanoje, S. Girase, and D. Mukhopadhyay, 'User profiling trends, techniques and applications', ArXiv Prepr. ArXiv150307474, 2015.
[13] H. Kim and P. K. Chan, 'Implicit indicators for interesting web pages', 2005.
[14] J. Xiao, Y. Zhang, X. Jia, and T. Li, 'Measuring similarity of interests for clustering Web- users', in Proceedings 12th Australasian Database Conference. ADC 2001, 2001, pp. 107–114.
[15] H. Liu and V. Kešelj, 'Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users' Future Requests', Data Knowl Eng, vol. 61, no. 2, pp. 304–330, May 2007.
[16] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, 'A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites', IEEE Trans. Knowl. Data Eng., vol. 20, no. 2, pp. 202–215, Feb. 2008.
[17] T. Htwe and N. S. M. Kham, 'Extracting data region in web page by removing noise using DOM and neural network', in 3rd International Conference on Information and Financial Engineering, 2011.
[18] R. P. Velloso and C. F. Dorneles, 'Automatic Web Page Segmentation and Noise Removal for Structured Extraction using Tag Path Sequences', J. Inf. Data Manag., vol. 4, no. 3, p. 173, Sep. 2013.
[19] Y. L. Sulastri, A. B. Ek, and L. L. Hakim, 'Developing Students' Interest by Using Weblog Learning', GSTF Int. J. Educ. Vol1 No2, vol. 1, no. 2, Nov. 2013.
[20] A. Nanda, R. Omanwar, and B. Deshpande, 'Implicitly Learning a User Interest Profile for Personalization of Web Search Using Collaborative Filtering', in 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014, vol. 2, pp. 54–62.
[21] J. Onyancha, V. Plekhanova, and D. Nelson, 'Noise Web Data Learning from a Web User Profile: Position Paper', in Proceedings of the World Congress on Engineering, 2017, vol. 2.

