**Web Traffic Forecasting Using ARIMA and LSTM**

Dr.A.K.MARIAPPAN1, S.V.NAGESWARI2, R.NIVEDHA NANDHINI3,

ANJANA SRIRAM4

1Professor, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. maris2612@yahoo.com  
2Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. [coolnacha99@gmail.com](mailto:coolnacha99@gmail.com)

3 Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. nivedhanandhini2000@gmail.com  
4 Student, Department of Information Technology, Easwari Engineering College, Chennai, Tamil Nadu, India. anjanasriram999@gmail.com

**Abstract** - In today’s world web traffic is one of the serious issues faced by many. Web traffic tends to hinder the smooth user experience and it is also very challenging for the web service providers to maintain a smooth user-server interaction. We are looking to overcome this problem by building a prediction model to forecast the web traffic in advance to avoid all the problems faced. Our model thoroughly studies the previous web traffic data to efficiently predict the web traffic of a particular website at a given point in time.

Forecasting is one of the important goals of mining time-series databases. The efficacy of Time series forecasting has been proved while decision making in various domains. This method is vastly different from the other proposed methods for prediction and analysis. This paper proposes the use of ARIMA and LSTM algorithms to forecast web traffic.

**Keywords** – Web traffic, Time series, ARIMA, LSTM, forecasting, Multivariate.

1. **INTRODUCTION**

The increase in web traffic could pose a hindrance to the workflow and also creates a lot of issues thus an organization is forced to find a way to manage the web traffic efficiently to be successful. Many would have come across extraordinarily slow websites that take an extendedtime to load once the traffic on their specific website is high, for instance, we may have gone over a few internet businesses sites that may crash when the number surpasses the expected limit. which causes a lot of weight for the customers and on account of that it could reduce the customer's evaluations of the site and rather use another site, thusly, diminishing their business. Appropriately, a traffic the board strategy or plan should be set up to diminish the peril of such occurrences which could be unfavourable to the presence of the association. Starting in the relatively recent past, there wasn't a necessity for such gadgets as most specialists could manage the traffic assembly anyway the phone age has extended the solicitation to an especially level for specific sites that associations couldn't have reacted quickly enough to keep up their standard customer support level.

Time arrangement is an information grouping in sequential request. There are numerous regular instances of time arrangement, as sun powered irradiance, probabilistic expectation, etc. Time arrangement research incorporates numerous viewpoints, for example, time arrangement investigation, information collection and capacity, time arrangement expectation. Time arrangement anticipating is to foresee future information or variety propensity by examining chronicled information and it is a hot exploration point due to its wide applications in account, development costs expectation, and some different fields.

Ordinarily, we have two well-known sorts of worldly information, first and the acclaimed one is time arrangement information, and the second is information with time focuses. Time arrangement information is a significant class of worldly information protests, an assortment of perceptions, which are in sequential request made. A period arrangement is a succession of noticed information, generally requested on schedule. Inside the composing various techniques have been proposed for expecting web traffic. They can be described completely into two social occasions reliant on the examined models: nonlinear assumption and direct conjecture. The most by and large used models Linear guess models are: I) HoltWinters Algorithm ii) AR Model iii) MA Model. The assessment focused on rehashing neural associations is consistently used for nonlinear gauge. Discrete wavelet change (DWT) parcels the data into straight and non-direct portions that help improve guess exactness. ES-RNN extends execution by using GPU figuring to set up the dataset.

Our proposed approach utilizes Long Short-Term Memory (LSTM) RNN. Adding a piece of new data to RNN changes the current data by adding a capacity. Subsequently, the entire data is refreshed, i.e., there is no regard for 'significant' data and 'not all that significant' data as a rule. Both RNNs have their present layer of input circles. It permits them to keep data and information in 'memory' extra time. In any case, it could be hard to prepare standard RNNs to take care of issues requiring long-haul fleeting conditions to comprehend. This is because the misfortune work slope rots dramatically extra time (called the issue of the disappearing inclination). LSTM networks are a sort of RNN that utilizes other than standard units, uncommon units. LSTM frameworks incorporate a 'memory cell' which can hold information in memory for significant periods. As data enters the memory when it is yielded and when it is lost a progression of doors is utilized to follow it. This design assists them with seeing longer-term conditions. GRUs are like LSTMs yet are fundamentally rearranged. They likewise utilize a progression of doors to control data streams, however don't utilize diverse memory cells and futile entryways. We use LSTM RNN for this impact to have more memory than traditional RNN.

1. **RELATED WORKS**

Rodrigo N. Calheiros et al. [1], they provide a Cloud based workload prediction module for SaaS suppliers using the Autoregressive Integrated Moving Average (ARIMA) model. They presented the prediction based on the ARIMA model and estimated its accuracy of future workload prediction by using the real traces of requests to the web servers. They also calculated the impact of the achieved accuracy with respect to the efficiency in resource utilization and QoS. Simulated results show that their model is able to accomplish an average accuracy of up to 91%, which further leads to efficiency in resource usage with minimum impact on the quality of service.

G., P. Zhang et al. [2], they present a hybrid methodology that consolidates both Autoregressive integrated moving average (ARIMA) and artificial neural networks (ANNs) models. The hybrid method takes advantage of the unique characteristics provided by ARIMA and ANN models in linear and nonlinear modelling. The hybrid methodology considers factors such as sampling variation, model uncertainty, and structure change to provide results thus experimental results with real datasets demonstrates that the consolidated model can be an effective method to improve forecasting accuracy accomplished by either of the models used individually.

Tejas Shelatkar et al. [3] they presented web traffic Time series prediction which is performed using Long Short-Term Memory Recurrent Neural Network (LSTM RNN) and Autoregressive integrated moving average (ARIMA) more efficiently and accurately. The system predicts the number of users who will access the website in the future. The system will keep on upgrading and produce accurate results as more user data is fed. The system can be used by any user for improving their web traffic load management and business analysis. LSTM RNN provides more accuracy to the system. The system effectively records seasonal and long-term patterns including information such as holidays, day of week, language, region which will help our model to capture the trends of the data more efficiently.

Saman Feghhi et al. [4] they introduced an attack on the encrypted web traffic that utilizes only the packet timing data on the uplink. This attack is therefore impenetrable to existing packet padding defences. Likewise, in contrast to existing approaches, this timing-only attack does not need the information on the start or end of the web fetches and so is effective against traffic streams. We exhibit the effectiveness of the attack against the wired and wireless traffic, accomplishing average success rates of 90%. Likewise, this timing-only attack serves to emphasize deficiencies in the already present defences and also to the areas where it would be useful for virtual private network (VPN) designers to concentrate their further attention.

Rishabh Madan et al. [5] they have presented a time series forecasting technique to forecast internet traffic based on prior values. Numerous forecasting techniques such as ARIMA are used for making predictions, but it is mostly convenient for a time series which is linear. Whereas, neural networks like RNN are capable of predicting time series which are nonlinear. The presented system uses Discrete Wavelet Transform (DWT) and uses a high pass filter and a low pass filter resulting in linear and nonlinear parts for the time series. The proposed technique is more efficient and accurate than ARIMA and RNN individually.

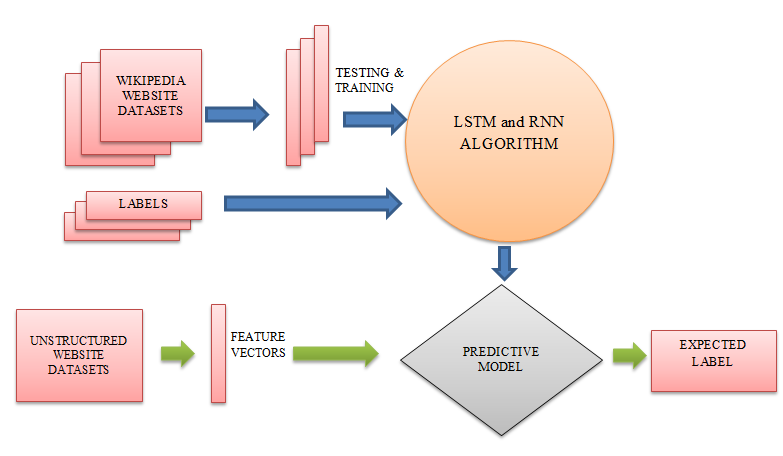
Navyasree Petluri et al. [6] they propose a system where they utilize existing Web Traffic Time Series Forecasting dataset by Google to forecast future traffic of the Wikipedia site. Forecasting web traffic is used to aids website owners to regulate an effective technique for load balancing of web pages present in the cloud, forecasting future patterns based on prior data and comprehend the user behaviour. They built a time-series model that uses RNN seq2seq model. They use symmetric mean absolute percentage error (SMAPE) for measuring the complete efficiency and accuracy of the developed model. Finally, evaluating the result of the developed model to the existing ones to determine the effectiveness of the presented method in forecasting future traffic of Wikipedia articles.

Seyyed Meysam et al. [7] they have proposed a system that deals with the issue of detecting DoS and DDoS attacks. Two features number of packets and source IP addresses are utilized as detection metrics that are calculated from network traffic every minute. Thus, a time series based on the number of packets is created using a Box-Cox transformation. An ARIMA model is also used for forecasting the number of packets in every minute. Then, using Lyapunov exponents and categorizing the chaotic behaviour the system differentiates normal traffic and attack traffics from one other. Simulation outcomes present that the proposed system can efficiently differentiate 99.5% of traffic states.

Soheila Mehrmolaei et al. [8] In this paper, they have proposed time series forecasting techniques to categorize and place two groups on the basis of forecasting duration. Moreover, a technique is presented by applying an average of estimation error for time series forecasting in ARIMA model. With respect to the outcome the improved ARIMA model is efficient than basic ARIMA model. As the future work, implementation of the application of the proposed approach in multivariate time series data sets can be implemented.

1. **PROPOSED SYSTEM**

At the point when the quantity of hits increments past the limit of a site, it will in general crash consequently making a gigantic misfortune for an organization. To evade this, we have come up with a prediction model which predicts the web traffic ahead of time with the goal that the necessary server can be assigned well ahead of time subsequently forestalling the event of an accident. This model aids in overseeing and limiting accidents adequately which thus forestalls the deficiency of an Organization. Our Model is a half and half multivariate model as our model is assembled utilizing both ARIMA and LSTM which radically builds the proficiency of individual calculations. ARIMA is best with linear data and LSTM is best with non-linear data. Our model is subsequently acceptable with the two kinds of reports. The yield of the ARIMA is given as input to the LSTM consequently training the dataset twice and thus acquiring better outcomes.



**Fig (1) System design**

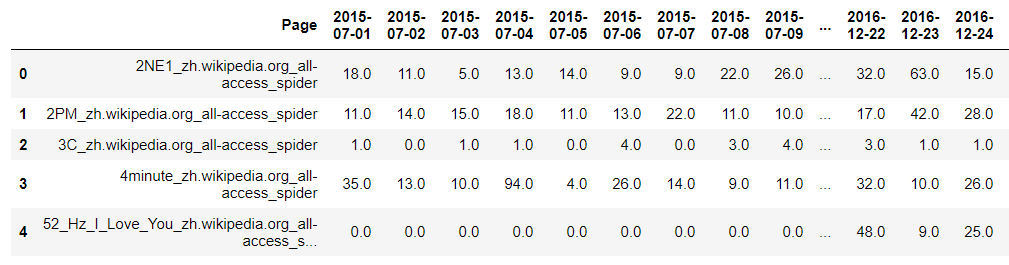
**Module wise implementation:**

* Pre-processing web traffic time series dataset.
* Training and validating with ARIMA
* Training and validating with LSTM-RNN
* Combining ARIMA and LSTM-RNN
* Deploying the forecasted model

**Pre-processing web traffic time series dataset:**

Data pre-processing is an information mining procedure that includes changing crude information into a reasonable configuration. Raw data is regularly deficient, conflicting, and additionally ailing in specific practices or drifts, and is probably going to contain numerous mistakes. Data pre-processing includes four steps, collecting, cleaning, storing and sharing.

As the first step, the dataset adopted for this project is daily views of Wikipedia articles provided by Kaggle comprising roughly 145,000 records. The dataset involves two fields, date and page. Page field comprises more than 1 lakh Wikipedia articles and the date field shows the number of hits day by day. The next step involves data cleaning, the gathered Wikipedia dataset may contain some missing values and this process fills the missing values with zero and organises the raw data for the following steps. When the dataset is cleaned and stacked, ensure that the dataset stored is right, prior to continuing further.



**Fig (2) Wikipedia Article**

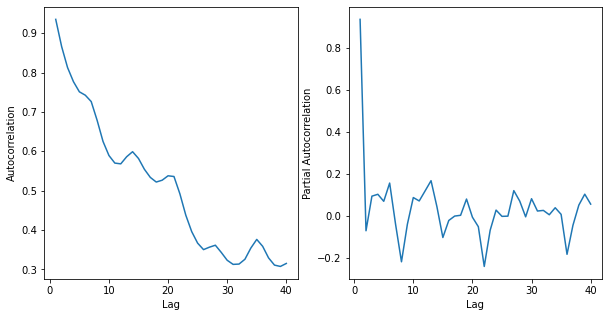
The refined dataset in its raw state cannot be supplied to the deep learning model which could lead to mis-leading prediction or an accuracy deficient model. To overcome these data-oriented challenges, the data set has to be segregated into a coarse-grained variant of itself which not only reduces the complexity of the dataset but also provides a clear and segregated view of data. This stage is called as data framing and in this stage, the dataset used to train the model is transposed from its existing state and classified based on the linguistic preferences of the user (hits based on user’s language preferences). Whence the dataset is segregated, the model is fed with the framed structure which initiates the training process. The model plots the graphical representation of the segregated dataset which serves as the pre-cursor for the whole prediction model which will be later put forth on the task of forecasting the future state of plausible user hits the target site / server could get at the predicted time frame.

**Training and validating using ARIMA:**

ARIMA (Auto-Regressive Integrated Moving Average) is really a class of models that 'clarifies' a given time arrangement dependent on its own previous values, that is, its own lags and the lagged forecast mistakes, so condition can be utilized to conjecture future values. ARIMA is isolated into three sections:

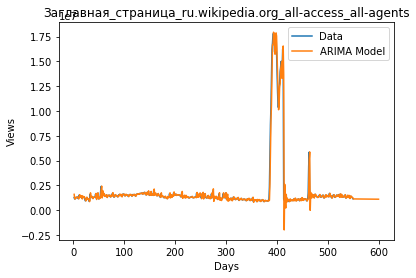
* Autoregressive (AR) forecasts future outcome from the past value.
* Integrated (I), it has to do with the distinction in time arrangement.
* Moving average (MA) model doesn't utilize the previous estimates to anticipate the future qualities though it utilizes the blunders from the past result.

Before training the data frames ARIMA model goes through different steps and one of them is it plot autocorrelation function (ACF) and partial autocorrelation (PACF) to distinguish the potential MA and AR model.



**Fig (3) ACF and PACF**

Based on the ACF and PACF values best ARIMA fir model is found for training and validating. In this manner, the forecast outcome is generated and plotted with best model.



**Fig (4) ARIMA model**

ARIMA model at times can anticipate the week-by-week base of the sign, which is acceptable. In different cases, it appears to simply give a direct fit. This is conceivably valuable.

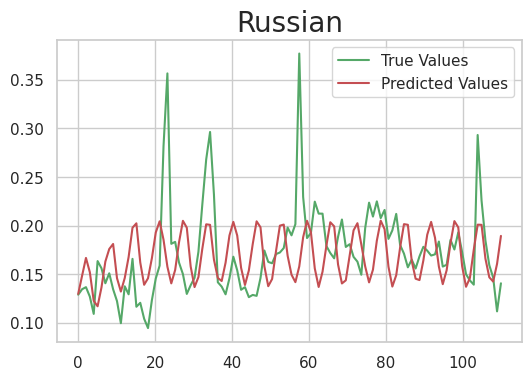
Be that as it may, on the off chance that we just aimlessly apply the ARIMA model to the entire dataset, the outcomes are not close to the same as utilizing the basic models. It actually appears to make them interesting properties, so perhaps we can consolidate this with another model to improve results.

The performance of LSTM is evaluated by Mean Absolute Percentage Error (MAPE) is 15.704892.

**Training and validating using LSTM:**

LSTM represents Long short-term memory is a self-supervised learning method, it is appropriate for both univariate and multivariant dataset. In this project, the multivariate dataset is utilized.Amultivariate dataset implies where there is more than one field to forecast.

Subsequent to preprocessing, the dataset is split for training and testing. Converting the split dataset into NumPy array and reshaped the array (3D) to which the LSTM model accepts. At that point construct the LSTM design. Fabricated model train and test the dataset for evaluating the performance.

****

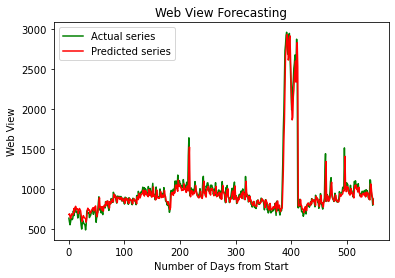
**Fig (5) LSTM model**

The performance of LSTM is evaluated by Mean Absolute Percentage Error (MAPE) is 35.686939.

**Combining ARIMA and LSTM:**

The over two models are fitted dependent on their best. As ARIMA model suits well for the linear dataset and LSTM works out positively for the non-linear dataset. As an individual outcome, both show their anticipated worth with some flaws. For better precision and result, the two models are ensembled together.

As the blend of both the models, from the outset ARIMA model's yield is given as the contribution of the LSTM model. Along these lines the dataset is trained twice. With this mix, the precision level expanded and the rate blunder is decreased. MAPE score for the ensembled model is 7.862425, lower than ARIMA and LSTM MAPE score exclusively.

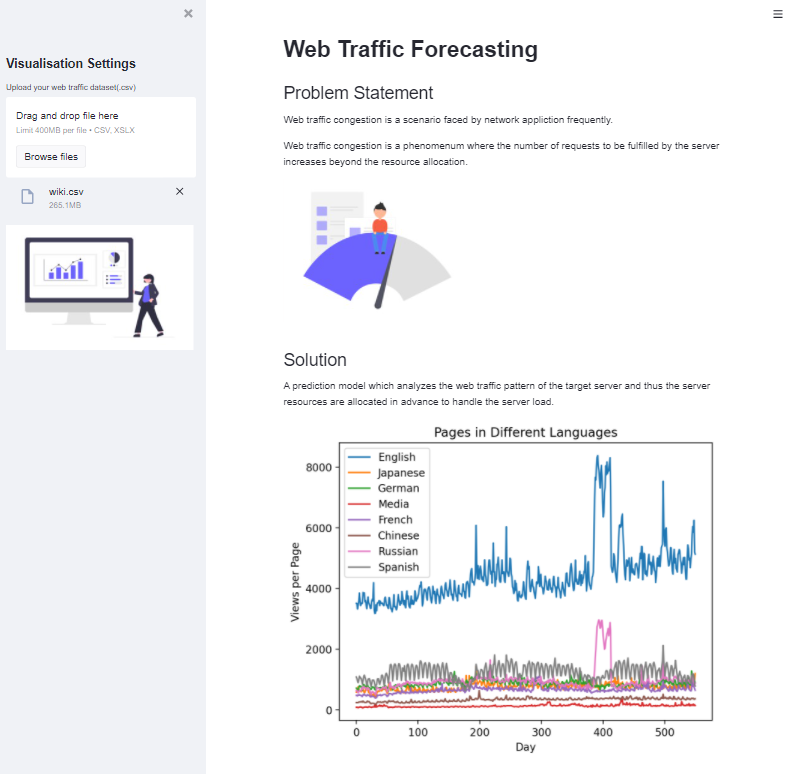


**Fig (6) Ensemble Model**

**Deploying the forecasted model:**

When the models are prepared and approved with ARIMA and LSTM, the forecasted model is incorporated with a website. The application is created with streamlit which is an open-source structure for deploying ML models.

The customer can discover their site traffic by uploading the past hits (ensure it is a multivariate dataset). The customer can see their traffic in graphical portrayal and can download the traffic record as a text file.



**Fig (7) Website**

1. **EXPERIMENTAL RESULT**

The performance evaluation for the time series forecasting model is calculated with the underneath methods,

1. Mean Forecast Error (MFE)
2. Mean Absolute Error (MAE)
3. Mean Squared Error (MSE)
4. Root Mean Squared Error (RMSE)
5. Mean Absolute Percentage Error (MAPE)

|  |  |  |  |
| --- | --- | --- | --- |
| **Methods** | **ARIMA** | **LSTM** | **Hybrid** |
| MFE | 0.000204 | 0.01451 | 0.00563 |
| MAE | 0.188291 | 0.09685 | 0.08114 |
| MSE | 0.054247 | 0.02006 | 0.02240 |
| RMSE | 0.232909 | 0.14164 | 0.11493 |
| MAPE | 15.704892 | 35.6869 | 7.86242 |

**TABLE I: Error Metrics**

1. **CONCLUSION**

Web traffic is a major issue these days. It makes sites crash impeding the smooth client experience, subsequently making a difficult issue for the organization. To take care of this issue, A web traffic forecasting model is fabricated utilizing ARIMA and LSTM, which proficiently predicts the web traffic in advance, and thereby the server can be allocated based on the requirement and numerous issues identified with web traffic can be addressed. We have created a website whereby uploading the previously obtained traffic as a CSV file, the website predicts the web traffic.

1. **FUTURE WORKS**

We have created a website whereby uploading the previously obtained traffic as a CSV file, the website predicts the web traffic. Future work would be to implement this as a plug-in, by using it, the service provider can get the predicted traffic in an instant.

**References:**

1.“Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications’ QoS” by Rodrigo N. Calheiros, Enayat Masoumi, Rajiv Ranjan, Rajkumar Buyya, 2014.

2."Time series forecasting using a hybrid ARIMA and neural network model", by G. P. Zhang, 2003.

3. “Web Traffic Time Series Forecasting using ARIMA and LSTM RNN” by Tejas Shelatkar, Stephen Tondale, Swaraj Yadav, Sheetal Ahir, 2020.

4. “A Web Traffic Analysis Attack Using Only Timing Information” by Saman Feghhi and Douglas J. Leith, August 2016.

*5.* “Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA and RNN” by Rishabh Madan, Partha Sarathi Mangipudi.

6. “Web Traffic Prediction of Wikipedia Pages” by Navyasree Petluri, Eyhab Al-Masri, 2019.

7. “A Novel DoS and DDoS Attacks Detection Algorithm Using ARIMA Time Series Model and Chaotic System in Computer Networks” by Seyyed Meysam Tabatabaie Nezhad, Mahboubeh Nazariy, and Ebrahim A. Gharavo, 2015.

8. “Time series forecasting using improved ARIMA” by Soheila Mehrmolaei,2016