Web Traffic Time Series Forecasting using ARIMA and LSTM RNN

Tejas Shelatkar[1,*], *Stephen* Tondale[2,**], *Swaraj* Yadav[3,***], and *Sheetal* Ahir[4,****]

[1]Ramrao Adik Institute of Technology, Nerul
[2]Ramrao Adik Institute of Technology, Nerul
[3]Ramrao Adik Institute of Technology, Nerul
[4]Ramrao Adik Institute of Technology, Nerul

Abstract.
Nowadays, web traffic forecasting is a major problem as this can cause setbacks to the workings of major websites. Time-series forecasting has been a hot topic for research. Predicting future time series values is one of the most difficult problems in the industry. The time series field encompasses many different issues, ranging from inference and analysis to forecasting and classification. Forecasting the network traffic and displaying it in a dashboard that updates in real-time would be the most efficient way to convey the information. Creating a Dashboard would help in monitoring and analyzing real-time data. In this day and age, we are too dependent on Google server but if we want to host a server for large users we could have predicted the number of users from previous years to avoid server breakdown. Time Series forecasting is crucial to multiple domains.
ARIMA; LSTM RNN; web traffic; prediction;time series;

1 Introduction

Recently, more and more people are getting access to the internet all over the world, the rise in traffic for almost all websites are inevitable. The increase in traffic for the websites could cause a lot of problems and the company which manages to cope with the traffic changes in the most efficient way is going to succeed.[7] As most of the people may have encountered a crashed site or very slow loading time for a website when there are a a lot of people using it, like when various shopping websites may crash just before festivals as more people try to log into the website than it was originally capable of which causes a a lot of inconveniences for the users and as a result of that it could decrease the user's ratings of the site and instead use another site, therefore, reducing their business. Therefore, a traffic management technique or plan should be put in place to reduce the risk of such mishaps which could be detrimental to the existence of the company. Until recently, there wasn't a need for such tools as most servers could handle the traffic influx but the smartphone age has increased the demand to such a high level for some websites that companies could not have reacted quickly enough to maintain their regular customer service level.
Within the literature many methods have been proposed for forecasting web traffic. They can be classified broadly into two groups based on the analysed models: nonlinear prediction and linear prediction. The most widely used models Linear forecast models are: i) HoltWinters Algorithm ii) AR Model iii) MA Model. The forecasting focused on recurring neural networks is commonly used for nonlinear prediction. Discrete wavelet transform (DWT) divides the data into linear and non-linear components that help improve forecast accuracy.[1]
ES-RNN increases performance by using GPU computing to train the dataset.

2 Literature Survey

Here, we explain the existing technology in the field of web traffic time series forecasting and also the data set that has been used for our prediction model.

2.1 Web Traffic Time Series Dataset

Wikipedia's pageview API is the data used for this project.That data contains daily page visits as a time series to any post. Latest data is obtained through this API. The data is returned in JSON format. The fields extracted from this data are the recorded Dates and Visits on that date. It converts this data into a data frame and fits into the predictive model. [7,8]

2.2 DWT

Discrete Wavelet Transform is used to divide the data vector by two sections into precise coefficients and approximate coefficient. A sequence of low pass filter and high pass filter are used to evaluate the DWT. This filtering produces Data Thorough and Approximate, a low frequency component, and a high frequency. After this step the two

[*]e-mail: iamtejasshelatkar@gmail.com
[**]e-mail: stephentondale7@gmail.com
[***]e-mail: swarajy26@gmail.com
[****]e-mail: sheetal.ahir@rait.ac.in

components are reconstructed by passing D and A through the inverse discrete wavelet transformation (iDWT).[1]

2.3 ARIMA MODEL

ARIMA (Auto-Regressive Integrated Moving Average) the model has a huge advantage in univariate time series forecasting. ARIMA model attempts to describe the trends and seasonality in time series as a function of lagged values(Auto Regressive parameter) and Averages changing over time intervals(Moving Averages). The model includes differencing (Integrating) the original time series data. Differencing time-series means forming a new time series by subtracting the previous observation from the current time. The point of this is to remove certain trends, such as seasonality, trends, or inconsistent variance in time series data.[5] The ARIMA equation has two important components Auto-Regressive (AR) part and the Moving Average (MA) part.

2.4 LSTM RNN

Our proposed methodology uses Long Short Term Memory (LSTM) RNN.To add a piece of new information to RNN, it completely transforms the existing information by adding a function. As a result, the whole information is updated, i.e. there is no respect for ' important ' information and ' not so important ' information in general.

Both RNNs have the recurrent layer of feedback loops. It allows them to keep information and data in' memory' over time. Nonetheless, it may be difficult to train standard RNNs to solve problems requiring long-term temporal dependencies to understand. This is because the loss function gradient decays exponentially over time (called the problem of the vanishing gradient). LSTM networks are a type of RNN that uses besides standard units, special units[3]. LSTM systems include a ' memory cell ' which can hold data in memory for long periods of time. As information enters the memory when it is output and when it is lost a series of gates is used to track it. This architecture helps them to understand longer-term dependencies. GRUs are similar to LSTMs but are structurally simplified. They also use a series of gates to control information flow, but do not use different memory cells, and use less gates. We use LSTM RNN for this effect to have more memory than conventional RNN.[2]

3 Problem Statement

The problem of forecasting the future values of time series has always been one of the most challenging problems in the field. Real time dashboard is a dashboard that contains visualizations that are automatically updated with the most current data available. These data visualizations offer a combination of historic data and real-time information that is useful for identifying emerging trends and monitoring efficiency. Real time dashboards usually contain data that is time-sensitive.

4 Proposed Methodology

Discrete wavelet Transform breaks down data signals into basic wavelet functions. Since the time-series data procured for investigation is noisy in nature it is very important to complete pre-processing of the data. Here the DWT splits down the data into pieces of low frequency and high frequency. So for best results, we can apply the algorithm to the segments. Since the data must be fixed in ARIMA henceforth we use high-frequency data as a predictive contribution. RNN uses data from low frequencies as input. It was later observed that this technique yields palatable results for less and more knowledge that is not the independently implemented situation for ARIMA and RNN.

In the following steps the proposed network traffic prediction methodology based on DWT, ARIMA, and LSTM RNN may be shown.

1. Load the time series for network traffic into a Xin vector.

2. Apply DWT to and decompose input variable. The DWT is determined via a series of low pass filter and high pass filter. This filtering produces data components of low frequency and high frequency called Detailed(D) and Approximate(A).

3. After this step, the two components are reconstructed via inverse discrete wavelet transformation (iDWT) by passing D and A.[1].

4. The ARIMA model is applied to the D component to produce forecast. BY compPicture1Residual Sum of Squares (RSS) of AR, MA, and ARIMA defines the vector-specific ARIMA model (p, d, q). Using ARIMA to estimate the linear component Yfor1 = forecast1('ARIMA model', Xd(eff)) / Xd(eff) is obtained by subtracting from the training set (p+d+q).

5. We use a Vanilla LSTM which has only a single hidden layer. We apply it to the A part. As input, forget, output get are fed then for cell state tanh activation function is used and also for the final output.

6. Combining forecasts from linear and nonlinear sections to get the final data forecast. This data can be compared with the actual data so we get errors for each algorithm.

,2 The system flow diagram is depicted in Figure 2. The figure indicates that the dataset will go through DWT to give output as linear and non-linear components. The ARIMA model will forecast for linear component and the LSTM-RNN will roll out forecasts for non-linear component. These forecasts will be combined using iDWT to provide a final forecast.

5 Implementation and Results

The dataset was analyzed and the sample data used was 'India'. Further, the dataset was divided into training and testing sets. For the time series, we plotted the number of hits vs. days along with real values and forecasts for the article 'India' during the testing period. The x-axis represents the Time Interval and the y-axis represents the page visits in powers of 10. The monthly forecast results

https://doi.org/10.1051/itmconf/20203203017

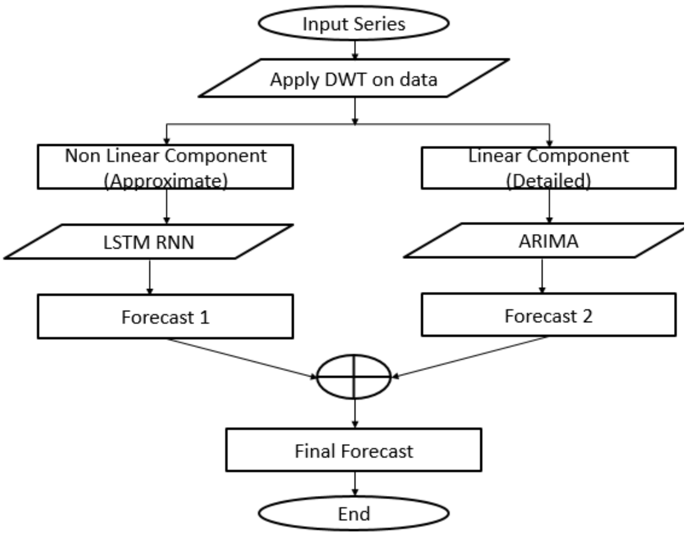

Figure 1. System Flowchart.

for languages obtained as per the proposed methods are as follows:

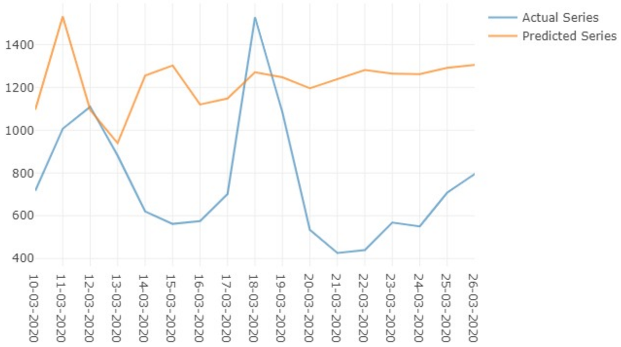

Figure 2. Forecast results for ARIMA

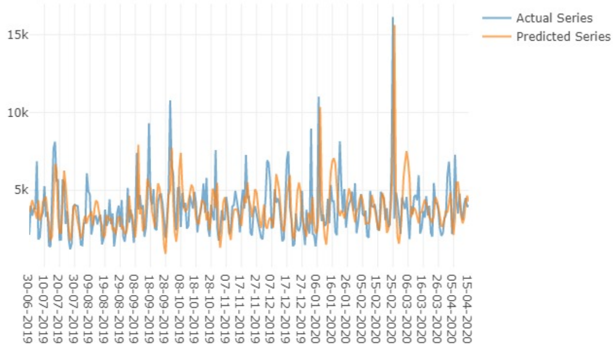

Figure 3. Forecast results for LSTM-RNN

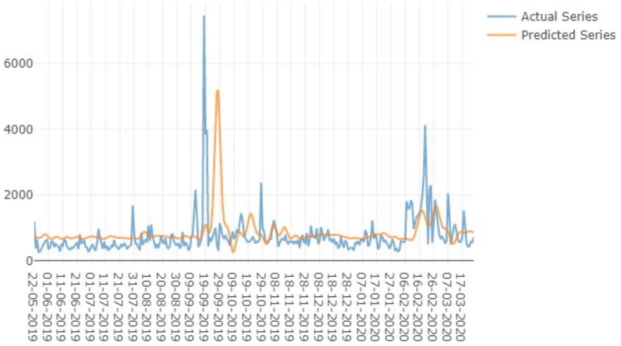

Figure 4. Forecast results after using DWT

The forecast results obtained in Figure 2 are for ARIMA model. There is a clear trend over the time interval and the forecast replicates the same. In Figure 3, the forecast results for RNN show a pattern which detects spikes accurately. Figure 4 depicts the use of DWT which combines results of ARIMA and RNN to provide more accurate results.

6 Conclusion

Web traffic Time series prediction can be carried out using Long Short Term Memory Recurrent Neural Network and Autoregressive integrated moving average more efficiently and accurately. Prediction of the number of users will access the website in the future is possible. The proposed will keep on improving as more user data is fed. Our system can be used across all websites for improving their web traffic load management and business analysis[5]. LSTM RNN brings more efficiency to our system. Our system effectively captures seasonal patterns and long-term trends Including information about holidays, day of week, language, region might help our model to capture more correctly the highs and lows.

7 Future Work

Time Series Forecasting is one of the least explored areas and various models are evaluated to improve the accuracy of the forecast. The main focus of the proposal is to predict future web traffic to make decisions for better congestion control. Past Values are considered to predict future values. We will also seek to explore multivariate time series and offer suggestions for simplifying the decision-making process in real-time.

References

[1] "Predicting Computer Network Traffic: A Time Series Forecasting Approach using DWT, ARIMA and RNN" by Rishabh Madan, 2018.
[2] “Fast ES-RNN: A GPU Implementation of the ES-RNN algorithm ” by Andrew Redd and Kaung Khin, 2019.
[3] “Time Series Forecasting Based on Complex Network Analysis“ by SHENGZHONG MAO AND FUYUAN XIAO, 2019.
[4] “Web Traffic Prediction of Wikipedia Pages” by Navyasree Petluri, Eyhab Al-Masri, 2019.
[5] “Time series forecasting using improved ARIMA” by Soheila Mehrmolaei,2016.
[6] “Efficient Prediction ofy Network Traffic for Real-Time Applications” by Muhammad Faisal Iqbal , Muhammad Zahid, Durdana Habib, and Lizy Kurian John, 2019.
[7] https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews
[8] https://towardsdatascience.com/3-facts-about-time-series-forecasting-that-surprise-experienced-machine-learning-practitioners-69c18ee89387.
[9] "Temporal Pattern Attention for Multivariate Time Series Forecasting" by Shun-Yao Shih Fan-Keng Sun Hung-yi Lee, 2018.
[10]“Time series forecasting using improved ARIMA” by Soheila Mehrmolaei,2016.
[11] “Efficient Prediction of Network Traffic for Real-Time Applications” by Muhammad Faisal Iqbal , Muhammad Zahid, Durdana Habib, and Lizy Kurian John, 2019.
[12] “Modelling Approaches for Time Series Forecasting and Anomaly Detection” by Shuyang Du , Madhulima Pandey, and Cuiqun Xing, 2018.
[13] “Neural Decomposition of Time-Series Data for Effective Generalization” by Luke B. Godfrey and Michael S. Gashler, 2017