Web Traffic Time Series Forecasting using Machine Learning

Assiya Muskan, Ashwini, Chinmayee R, Nayana R.S, Girish L
Department of Computer Science & Engineering, Visveswaraya Technological University, Channabasaveshwara Institute Of Technology, Gubbi, Tumkur, Karnataka, India
assiyamuskan@gmail.com

ABSTRACT:

Now days, web traffic anticipating is a significant issue as this can make misfortunes the activities of major sites. Time - arrangement topics has been an interesting issue for research. Anticipating future time arrangement esteems is one of the most troublesome issues in the business. The time arrangement field includes various issues, running from induction and examination to gauging and grouping. Estimating the organization traffic and showing it in a dashboard that updates continuously would be the most productive approach to pass on the data. Making a dashboard would help in checking and dissecting continuous information. These days, we are excessively reliant on Google worker however in the event that we need to have a worker for huge clients we might have anticipated the quantity of clients from earlier years to stay away from worker breakdown. Time Arrangement anticipating is significant to various areas. These days, web traffic anticipating is a significant issue as this can make misfortunes the activities of major sites. Time-arrangement gauging has been an interesting issue for research. Anticipating future time arrangement esteems is one of the most troublesome issues in the business. The time arrangement field includes various issues, running from induction and examination to gauging and grouping. Estimating the organization traffic and showing it in a dashboard that updates continuously would be the most productive approach to pass on the data. Making a Dashboard would help in checking and dissecting continuous information. These days, we are excessively reliant on Google worker however in the event that we need to have a worker for huge clients we might have anticipated the quantity of clients from earlier years to stay away from worker breakdown. Time Arrangement anticipating is significant to various areas.

I. INTRODUCTION:

It is fundamental for information researchers and business investigator to acquire time arrangement insightful abilities. Time arrangement data set had been the quickest developing class of data sets in the previous two years, and both customary ventures and arising innovation businesses had been creating additional time arrangement information. A few instances of time arrangement information bases are the monetary market data set, climate estimating data set, smart home monitoring database, and supply chain monitoring database.

It is essential for data analysts and business agent to obtain time course of action keen capacities. Time plan informational collection had been the speediest creating class of informational indexes in the past two years, and both standard endeavors and emerging development organizations had been making extra time course of action data. A couple of

examples of time course of action data bases are the money related market informational index, environment assessing informational collection, sharp home checking informational collection, and stock organization noticing informational index.

Now days, more and more people are getting access to the internet all over the world, the rise in traffic for almost all websites are unavoidable. The increment in traffic for the websites could cause a lot of disputes and the company which survives to management with the traffic changes in the most systematic way is proceed to succeed. As most of the people may have very slow loading time for a website when there are a lot of people using it, like when various shopping websites may pound just before commemoration as more people try to log into the website than it was originally efficient of which just a good deal of disruption for the users and as a result of that it could downturn the user's ratings of the site and alternatively use another site, consequently, shorten their business. Accordingly, a traffic management technique or plan should be put in place to lessen the risk of such allotment which could be damaging to the existence of the company. Until latterly, there wasn't an essential for such tools as most servers could handle the traffic inscription but the smartphone age has enlarged the ultimatum to such a high level for some websites that companies could not have reacted immediately enough to continue their orderly customer service level. Evaluating web traffic on a web server is highly essential for web service providers since, without a conventional dictate forecast, customers could have lengthy bide one's time and spontaneity that website. Nevertheless, this is a backbreaker task since it essential to make dependable predictions based on the arbitrary nature of human behavior. We bring out an architecture that gathered source data and in a supervised way executes the forecasting of the time series of the page views.

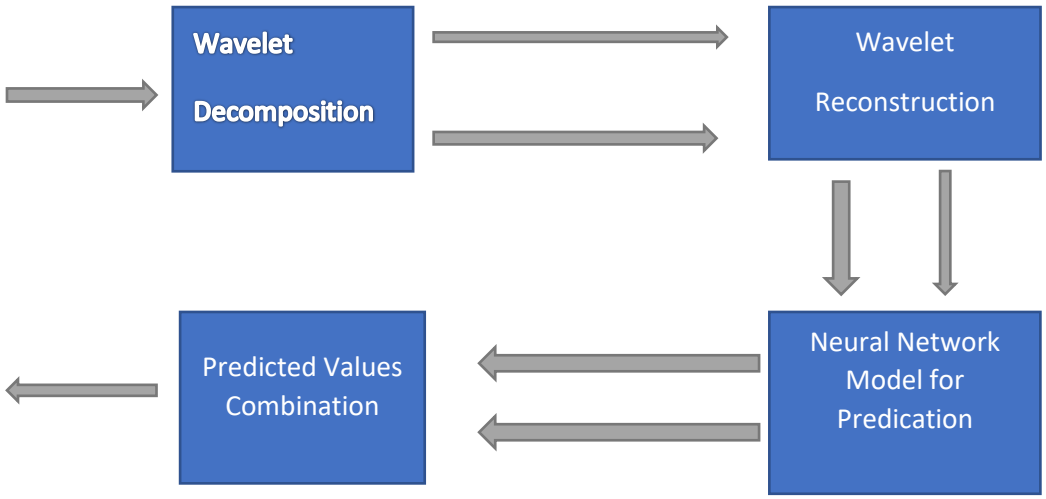
There are various researches going on in this space of gauging web traffic. As a context oriented examination up to now ARIMA, Holt winter likewise, various standard procedures are used at this point for better result the neural association comes into area with wavelet plan assessment for predicting web traffic. Though this is direct and supportive system for figures of web traffic prong game plan, it is sensibly jumbled similarly as drawn-out expressly for constant assessment of web traffic and its adaptability all through a gigantic time game plan data issue. Another creator had a go at utilizing the triumphant model of RNN (Repetitive neural organization) seq2seq model with middle as an extra component in various time periods, it could have been additionally improved with more examination on upgraded model and highlight set. There have been colossal number of exploration endeavors in contemplating and estimating web traffic and least examination works are zeroing in on the time-related perceptions which are arising for the examination and obstruction to arrangement and figure in assisting with assessing future visits of site pages. Thus, in this exploration we are intending to utilize some more list of capabilities to take care of the model with worldly and occasional spikes in web traffic and further improve the model design to estimate future traffic to pages all the more precisely and increment the dependability of forecasts results.

The remainder of the paper is divided as follows. Section 2, the related work of web traffic time series. Section 3, literature survey with model. Section 4, Results and implementation. Section 5. Finally , the conclusion with future work of web traffic time series forecasting.

II. RELATED WORK:

After customary strategies proposed model is wavelet design examination utilized with neural organization investigation of page visits and BP (Back Engendering) neural organizations for expectation of the site traffic. In this strategy wavelet change is applied to time arrangement information of unique site traffic and they noticed one low-recurrence development signal and a few high-recurrence detail signals are created. They at that point apply BP Neural Organization forecast model to each flag created lastly the anticipated aftereffects of all signs are joined into definite traffic anticipated worth which in outcome is an expectation result for a solitary page on specific day.

Web traffic time series is decomposed by the wavelet function to different signals and get high-frequency signals and one of the levels as low-frequency signal[1]. Then they will reconstruct each branch with signal branches that match the same length as time series. They use BP Neural Network models for forecasting each signal they have tested the algorithm on a small dataset of 61 days. The model was trained with first 60 days traffic as training data and traffic on day 61 was used as test data



After the wavelet based neural network RNN Seq2Seq model was proposed on Wikipedia's web traffic time series data and analysed total count of visitors per article of Wikipedia, number of visits on articles, features decomposed from page URLs, weekly & year-year seasonality information, lagged page views and page popularity[2]. For improvement in RNN Seq2Seq rebuilt the existing RNN Seq2Seq model using encoder-decoder. The results generated from the decoder were used as inputs for the next step till the end of the sequence for a particular batch size. This encoder-decoder performs better. They include median approach as the most important step in dataset behaviour. They used two main features in existing model rolling median and Fibonacci median, which is slight improvement in existing model.

In a web traffic there are might be unexpected spikes in visits which is issue for improving outcomes. Along these lines, to affect less precision to fate of same, they utilized moving middle as control to keep away from high effect for the progressive sections of information [3]. For patterns they utilized moving window to gauge get fame of articles dependent on

week by week, month to month, quarterly and yearly moving window size is taken as 7, 30, 90, 180 days as moving window.

The issue in the Wikipedia time series of estimating the future upsides of time arrangement has consistently been perhaps the most difficult issues in the field. Continuous dashboard is a dashboard that contains perceptions that are naturally refreshed with the most current information accessible. These information perceptions offer a mix of memorable information and continuous data that is helpful for recognizing arising patterns and observing proficiency. Continuous dashboards ordinarily contain information that is time-delicate which is shown in the implementation and result.

III. 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.

ARIMA MODEL

ARIMA represents auto-backward incorporated moving normal. It is perhaps the most widely recognized and solid models utilized in Time Arrangement expectations. It contains Auto regression (AR), Coordinated (I) and Moving-normal (Mama). Auto regression is "A model that utilizes the reliant connection between a perception and some number of slacked perceptions."

Incorporated is "a model that utilizes the differencing of crude perceptions (for example deducting a perception from the past time step). Differencing in measurements is a change applied to time-arrangement information to make it fixed. “This permit the properties don't rely upon the hour of perception, dispensing with pattern and irregularity and settling the mean of the time arrangement."

Moving-normal is "a model that utilizes the reliance between a perception and a lingering mistake from a moving normal model applied to slacked perceptions. In spite of the AR model, the limited Mama model is consistently fixed."

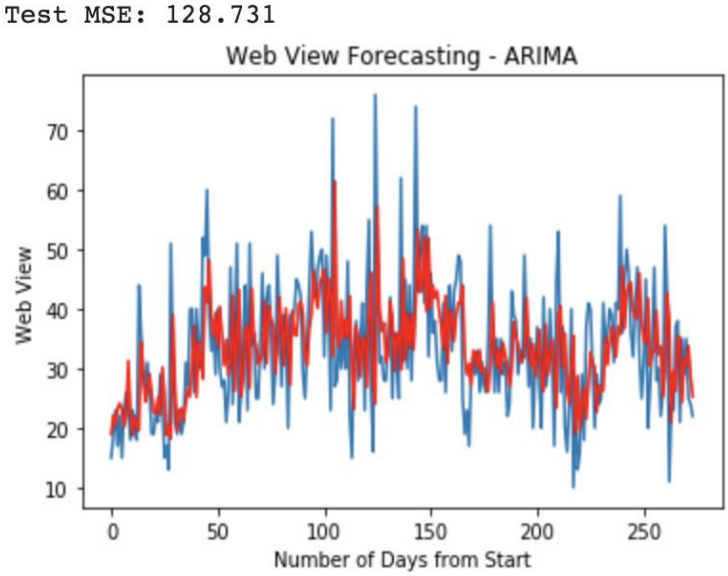
At the point when we utilize these three parts in the ARIMA model, it concocts three boundaries:

1. p (slack request): number of slack perceptions remembered for the model
2. d (level of differencing): number of times that the crude perceptions are differenced
3. q (request of moving normal): size of the moving normal window" .

BUILDING THE MODEL

The entire dataset from Wikipedia was enormous, and because of the computational intricacy, we zeroed in on Netflix Wikipedia page just and anticipated its example.

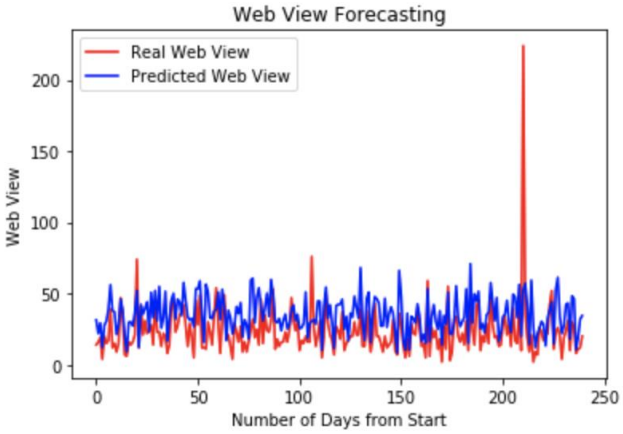
We discovered that ARIMA gives better outcomes with more modest datasets and we discovered that bunches of expectation focuses were not gathering their actual worth. There are different approaches to assess the model and discover suitable boundaries.



LSTM (LONG SHORT TERM MEMORY)

We have attempted the second strategy which is LSTM. It represents long momentary memory. It has a place with intermittent neural organizations. The manner in which we fabricate it is again to pick arbitrary line from the first dataset. We utilize a stage worth of 3 to rebuild the information and get our X and y into the displaying interaction. At that point we split the information into preparing and testing sets and use MinMaxScaler() to standardize the information. At long last, we reshape the information and train it.

1. Vanilla LSTM Model
2. Stacked LSTM Model
3. Bidirectional LSTM



LSTM RNN

Our proposed procedure utilizes Long Transient Memory (LSTM) RNN. To add a piece of new data to RNN, it totally changes the current data by adding a capacity. Accordingly, the entire data is refreshed, for example there is no regard for ' significant ' data and ' not all that significant ' data by and large. Both RNNs have the intermittent layer of criticism circles. It permits them to keep data and information in' memory' over the long haul. Regardless, it very well might be hard to prepare standard RNNs to tackle issues requiring long haul transient conditions to comprehend. LSTM networks are a kind of RNN that utilizations other than standard units, extraordinary units. LSTM frameworks incorporate a 'memory cell' which can hold information in memory for extensive stretches of time. This design assists them with seeing longer-term conditions. GRU's are like LSTMs yet are primarily improved. They likewise utilize a progression of doors to control data stream, yet don't utilize distinctive memory cells, and utilize less entryways. We use LSTM RNN for this impact to have more memory than traditional RNN.

Web Traffic Time Series Dataset

Wikipedia's site visit Programming interface is the information utilized for this venture. That information contains day by day page visits as a period arrangement to any post. Most recent information is gotten through this Programming interface. The information is returned in JSON design. The fields take out from this information are the enlisted Dates and Visits on that date. It changes over this information into an information casing and finds a way into the prescient model.

DWT

Discrete Wavelet Transform is used to split the data vector by two sections into recuse coefficients and approximate coefficient. A sequence of low pass filter and high pass filter are used to estimate the DWT. This filtering produces data thorough and approximate, a low frequency component, and a high frequency. After this step the two components are reconstructed by passing through the inverse discrete wavelet transformation (IDWT).

IV. PROPOSED METHODOLOGY

Discrete wavelet Transform (DWT) breaks down data signals into basic wavelet functions. Since the time-series data procedure for investigation is to complete pre-processing data. Here the Discrete wavelet Transform splits the data into two ways low frequency and high frequency. So for best results, we can apply the algorithm to the segments. Since the data fixed in ARIMA model hence we can use high-frequency data as a predictive contribution. LSTM 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 network traffic into data collection for verification.

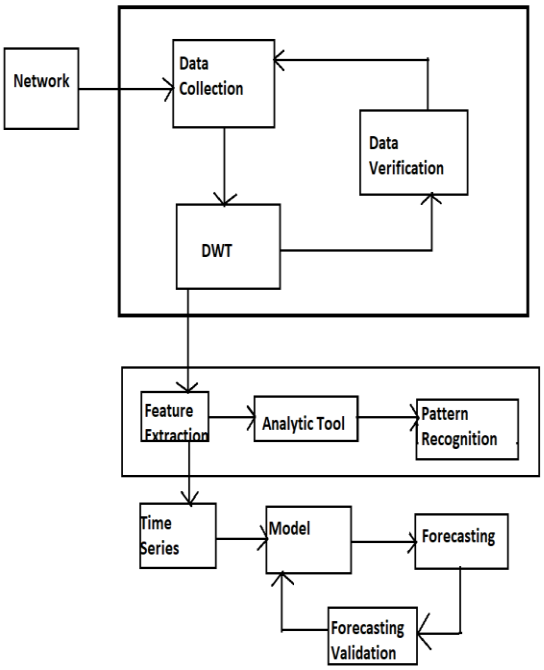
2. Apply DWT to decompose input variable. The DWT is determined via a series of low pass filter and high pass filter. This procedure to filtering the data components of low frequency and high frequency called as Detailed (D) and Approximate (A).

3. After this step of procedure, the two components are reconstructed is determined via of inverse discrete wavelet transformation (IDWT) by passing Detailed (D) and Approximate (A).

4. The ARIMA model is applied to the Detailed (D) component to produce forecast with remove certain trends, such as seasonality, trends, or inconsistent variance in time series data.

5. We use a Vanilla LSTM which has only a single hidden layer. We apply it to the Approximate (A) part to produce forecast.

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 in each algorithm.



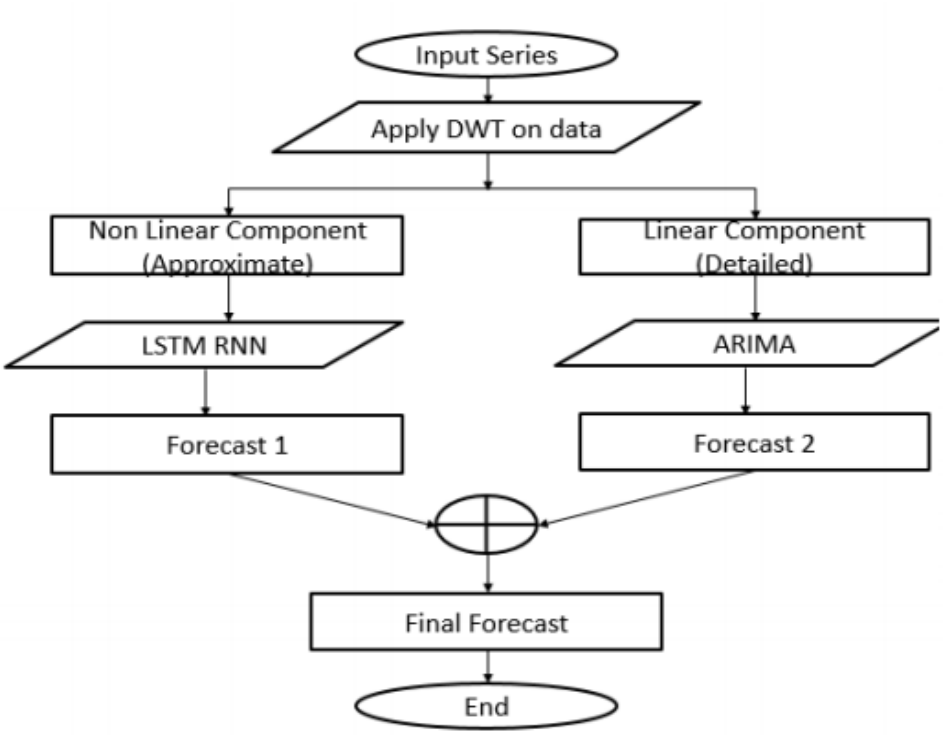


Figure 1. System Flowchart.

V. IMPLEMENTATION AND RESULTS

The dataset was analysed 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 result for languages obtained as per the proposed methods are as follows:

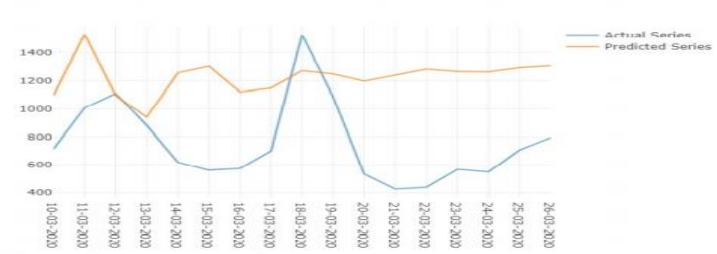


Figure 2. Forecast results for ARIMA

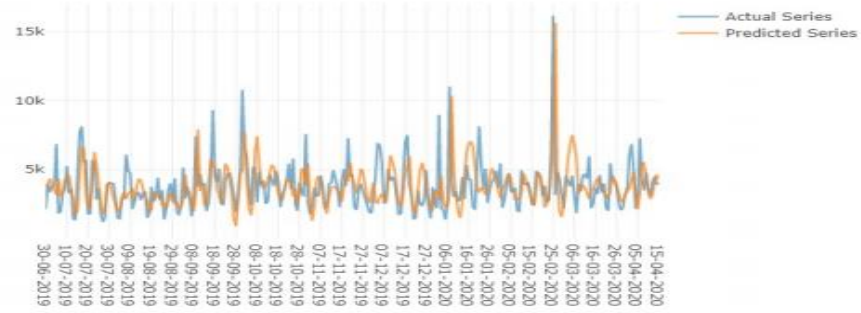


Figure 3. Forecast results for LSTM-RNN

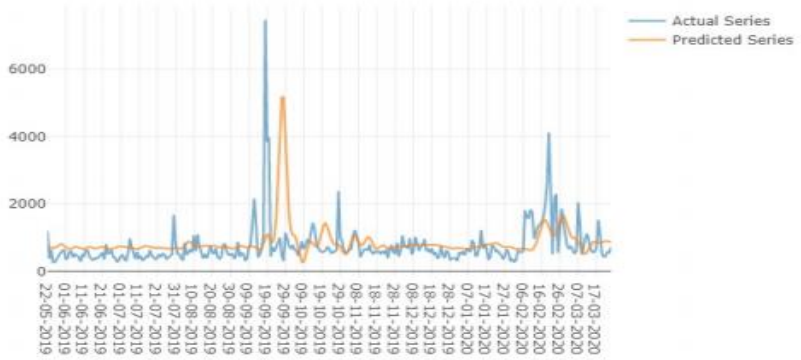


Figure 4. Forecast results after using DWT

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.

VI. Conclusion

Web traffic Time series prediction it can be achieved using Long Short Term Memory Recurrent Neural Network and Autoregressive integrated moving average more efficiently and accurately. How many number of users that will access the website/link in the future is may be predicted. The model that was proposed will keep on updating as many user data is fed. Our model can be played around all websites because of improving their web traffic load management and business analysis. More efficiency to our system can be get by LSTM RNN. Our system effectively captures seasonal patterns and long-term trends including information about holidays, day of week, language, and region might help our model to capture more correctly the highs and lows.

Time Series Forecasting is one of the least explored areas and various models are evaluated to improve the accuracy of the forecast. The main aim of our system is to predict future web traffic to make decisions for better congestion control. Previous data are considered to predict future values. We will also seems to explore various time series and provide a guidance for modulate 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] “Web Traffic Prediction of Wikipedia Pages” by Navyasree Petluri, Eyhab Al-Masri, 2019.

[3] “Time series forecasting using improved ARIMA” by Soheila Mehrmolaei,2016.

[4]“EfficientPredictionofNetworkTrafficforRealTimeApplications”byMuhammadFaisalIqbal, MuhammadZahid,DurdanaHabib,andLizyKurianJohn,2019.

[5]“ModellingApproachesforTimeSeriesFore-castingandAnomalyDetection”byShuyangDu,MadhulimaPandey,andCuiqunXing,2018.[13]“ NeuralDecompositionofTime-Series