An Approach to Predictive Analytics of Website Visitors Traffic and Pageviews

Sneha Gupta
Department of Computer Science & Engineering
Priyadarshini Bhagwati College of Engineering, PBCOE
Nagpur, India
snehathebest90916@gmail.com

Manoj S. Chaudhari
Department of Computer Science & Engineering
Priyadarshini Bhagwati College of Engineering, PBCOE
Nagpur, India
manojchaudhary2@gmail.com

***Abstract*—Ability of predicting website visiting patterns has a significance value for every website owner for targeting their business for right customers at right time. The available solutions to predict future access and usage patterns for websites are available as Business Intelligence tools. But many small and medium sized companies or website owners are unable to afford them since they are very expensive solutions. The freely available Business Intelligence tools such as Google Analytic doesn't support forecasting feature. Through this paper we will try to identify methodology to detect the trends and seasonal patterns from page view time series data and identify dominant factors in the variation of such series. We will also investigate the various procedures and a state space model for making relatively short-term prediction.**

***Index Terms*—Time series analysis, website pageview prediction, linear regression, smoothing techniques.**

I. INTRODUCTION

In the last few years, Internet has become the integral part of our day-to-day life. A growing number of applications in diverse areas such as personal communications, entertainment, education, telemedicine and defence rely on the internet to transmit the information. With the technological changes the accessibility and uses of the website in day to day life has increased tremendously which resulted in the large numbers of the competitors in all the sectors. The ability of predicting website visiting patterns has a significance value for any website owner for targeting their business for right customers at right time. The available solutions to predict future access and usage patterns for websites are available as Business Intelligence tools. But small and medium sized companies or website owners are unable to afford them since they are very expensive solutions. The freely available Business Intelligence tool such as Google Analytic doesn't support forecasting features. They are focused on analyzing user behaviours on the website and log related results. For instance in software products related companies, they need to know the visiting history for their sites to plan their new releases, upgrades, etc. Another case is density of forecast visitors would be helpful to allocate or de-allocates servers [1]. There will be a number of benefits of predicting the website visitor behavior. The following are few of them [2]:

1) *Prediction may add value in both an agency and in-house setting*: It will provides a more accurate way to set goals and plan for the future, which can be applied to client projects, internal projects, or overall team or department strategy.
2) *Prediction will help to create accountability for your team*: It will allow the website owner to continually set goals based on projections and monitor performance through predicted data accuracy.
3) *Predication may teach about inefficiencies in your team, process, and strategy*: The more we can segment the website prediction, the deeper we can dive into finding the root of the inaccuracies in the business projections. And the more granular we will get the data, the more accurate we can forecast.
4) *Predication is money*: It's help to market the product more efficiently and helps in cost effective marketing.

The fact that one can improve inefficiency in the website marketing process and strategy through forecasting means we can effectively increase ROI. Every hour and resource allocated to a strategy that doesn't deliver results can be reallocated to something that proves to be a more stable source of increased organic traffic. So finding out what strategies consistently deliver the results the website owner expect, means they will invest money into resources that have a higher probability of delivering a larger ROI. Furthermore, providing accurate projections gives the reviewer a more compelling reason to invest in the work that backs the forecast.

II. DEFINATION OF PREDICTION

Prediction is analyzing the future behavior of the selected data set. It is the process of estimating a future event based on recent and past time series data. It may not reduce the uncertainty of future however, it gives the decision makers an idea and a basic premise for planning [1].

A. What is business intelligence data?

Business intelligence is a decision support system where information is gathered for the purpose of predictive analysis and support for business decisions.

B. What is predictive analytics?

Predictive analytics is using business intelligence data for forecasting and modeling. It is a way to use predictive analysis data to predict future patterns.

C. How data mining helps predictive analysis?

Data mining aids predictive analysis by providing a record of the past that can be analyzed and used to predict which customers are most likely to renew, purchase or purchase related products and services.

Predictive analytics can aid in choosing marketing methods, and marketing more efficiently. By only targeting customers who are likely to respond positively, and targeting them with a combination of goods and services they are likely to enjoy, marketing methods become more efficient. In the best cases, predictive analytics can reduce the amount of dollars spent to close a sale. At its most effective, business intelligence data mining can help marketing professionals anticipate and prepare for customer needs, rather than just reacting to them. And data mining can present data on demographics which may have been previously overlooked. When applied to marketing strategy, predictive analytics and data mining can help managers to bring in more sales, while spending less on campaigns.

Businesses around the globe have been predicting their sales for a long time. The primary reason for it is planning, for instance, amount of inventory to store or when to allot the budget on marketing campaign etc. Now a day every e-Commerce business, big or small, has their online presence. This gives them an opportunity to accumulate data about their consumer's behavior, demographics, sources, number of new visits, etc, which can be indirectly used to predict sales. A precursor to sales can also be found by calculating some correlation between the number of visitors and sales. Since many consumers thoroughly research the products and services online before they buy, the web analytics predicted number of visitors will quickly alert you on any new trend, than what the sales data can.

The process of website traffic prediction can be implemented using the time-series approach and decomposing the website page views. People may visit a particular website for many different reasons which are next to impossible for us to fathom all the underlying factors. So, we will presume to know nothing about the causality that affects the variable we are trying to predict. Instead, we will examine the past behavior of a time series in order to infer something about its future behavior. Time-series models are particularly useful when little is known about the underlying process one is trying to predict.

III. DEFINATION OF TIME SERIES

"A time series is a set of observation taken at specified times, usually at equal intervals".
"A time series may be defined as a collection of reading belonging to different time periods of some economic or composite variables". **by –Ya-Lun-Chau** [20]

A. Components of a Time series [21]

Secular Trend(T): *Gradual long term movement(up or down). Easiest to detect. e.g: Population growth In India.*

Cyclical Patterns(C): Results from events recurrent but not periodic in nature. An up-and-down repetitive movement in demand repeats itself over a long period of time. e.g. Recession in US Economy.

Seasonal Pattern(S): Results from events that are periodic and recurrent in nature. An up-and-down repetitive movement within a trend occurring periodically. Often weather related but could be daily or weekly occurrence. e.g. Sales in festive seasons.

Irregular Component (I): Disturbances or residual variation that remain after all the other behaviors have been accounted for. Erratic movements that is not predictable because they do not follow a pattern. e.g. Earthquake.

B. Building Time Series Model(TCSI):

The following equation helps us to build the time series model data.

$$O(t) = T(t) + S(t) + I(t) \text{ or } O(t) = T(t) * S(t) * I(t)$$

where:-
O(t) = observed series, T(t) = Trend component, S(t) = Seasonal , I(t) = Irregular component

1) *Step 1***:** Smooth the series & de trend the series.
2) *Step 2***:** Find out Seasonal component and adjust the data for seasonality.
3) *Step 3***:** See if there is still some trend/seasonality in the data & quantify it.

Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of cancelling the effect due to random variation. A widely used technique is ***smoothing***. This technique when properly applied reveals more clearly the underlying trend, seasonal and cyclic components. Smoothing techniques are used to reduce irregularities in time series data. Smoothing techniques, such as the Moving Average, Weighted Moving Average, and Exponential Smoothing, are well suited for one-period-ahead prediction.

1) ***Moving Average***: In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean (MM) or rolling mean and is a type of finite impulse response filter.
2) ***Weighted Moving Average***: It is any average that has multiplying factors to give different weights to data at different positions in the sample window. Mathematically, the moving average is the convolution of the datum points with a fixed weighting function. One application is removing pixelisation from a digital graphical image. In technical analysis of financial data, a weighted moving average (WMA) has the specific meaning of weights that decrease in arithmetical progression.

3) **Exponential smoothing**: It is a very popular scheme to produce a smoothed time series. It assigns exponentially decreasing weights as the observation gets older. In other words, recent observations are given relatively more weight in forecasting than the older observations.

IV. LITERATURE SURVEY

Prediction is analyzing and predicting the future behavior of the selected data set. In Predictive analysis knowledge from the analyzed data is used to predict future behaviors. It help in many ways in various domains such as controlling load balance, future marketing campaigns, allocating or de-allocating resources and caching, perfecting webpage's for improve performance. There are limited number of researches have been done on website related prediction. D. Ciobanu, C. E. Dinuca [5] done a research for predicting the next page to be visited by a web user. They have created a java program, using Net Beans IDE, which calculates the probability of visiting the pages using the page rank algorithm and counting links. The approach was using website log analysis to determine probabilities of visiting the pages. Their concept founded from the web-page ranking algorithm Page Rank. Taowei Wang, Yibo Ren completed a research on the suggesting methodology for personalized recommendation using collaborative filtering. Their system architecture and details on data preparation described in the research [6]. For improving the quality of personal recommendation, they have proposed a new personalized recommendation model which takes the good consideration of URL related analysis and combines the K-means algorithm. They have shown proposed model is effective and can enhance the performance of recommendation through results. Web Log Mining by an Improved AprioriAll algorithm research done by WANG tong HE Pi-lian shows that the possibility and importance on applying Data Mining technologies in web log mining and also emphasizes some problems in the conventional search engines. Further they offer an improved algorithm based on the original AprioriAll algorithm, which has been used in weblogs mining widely. Test results show the improved algorithm has a lower complexity of time and space [7]. Other than the above researches already done on web related predications it has been mentioned by many researchers that their future research areas will be focused on predicting area. Chandana Napagoda discussed about website visit forecasting using Data Mining techniques in 2013 [8]. They mainly concentrated on forecasting of web site visits by using prediction methods such as Gaussian, Linear regression, Multilayer Perceptron regression and SMO regression.

V. PREDICTION TECHNIQUES

The data present in the environment of website usage predication is time dependent series of data points. Modeling and explaining such data using statistical techniques is **'Time series analysis'**. The process of using a model to forecast future events based on known past events is 'Time series prediction'. Time series data such as website access usage data has a natural temporal ordering. Typically in data mining applications, each data point is an independent example of the concept to be learned and the ordering of data points within the set doesn't matter. But for time series data it is not the case. One approach of handling time series data is removing its temporal ordering, so that the standard propositional learning algorithms can process them. Here when removing the temporal ordering, the time dependent data should be encoded via additional input fields. These fields are known as 'lagged' variables. After data has been transformed, regression algorithms can be applied to learn a model. One approach is to apply multiple linear regressions. Also any method which is capable of predicting target can be applied. Lagged variables are the main mechanism by which the relationship between past and current values of a series can be captured by propositional learning algorithms [1].

A. Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University Of Waikato, New Zealand. Weka has an environment which is dedicated for time series analysis. It allows creating forecasting models, evaluating them and visualizing them. The approach of time series analysis in Weka is transforming the data into form that standard propositional learning algorithms can process. As above mentioned Weka does this by removing the temporal ordering of individual input examples by encoding the time dependent via additional input fields. Various other fields are also computed automatically to allow the algorithms to model [1].

B. Gaussian Process

In probability theory and statistics, Gaussian processes are a family of statistical distributions in which time plays a role. In a Gaussian process, every point in some input space is associated with a normally distributed random variable. Moreover, every finite collection of those random variables has a multivariate normal distribution. The distribution of a Gaussian process is the joint distribution of all those random variables, and as such, it is a distribution over functions [10].

C. Simple Linear Regression

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regressions. Confidence interval provides a useful way of assessing the quality of prediction.

In prediction by regression often one or more of the following constructions are of interest [9] [11][15]:

1) A confidence interval for a single future value of Y corresponding to a chosen value of X.
2) A confidence interval for a single point on the line.
3) A confidence region for the line as a whole.

D. Multiple Linear Regressions

It attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y [11][16].

E. Multilayer Perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable [12][15].

F. SMO Regression

Support vector classifier will be trained using polynomial or RBF kernels where John C. Platt's sequential minimal optimization algorithm is implemented here.

VI. CONCLUSION

Within the last few years, we have seen an explosive growth in web usability as a field. Given its infancy, it is not surprising that there are so few tools to assist web analysts. We presented a Time series approach with the regression technique for predicting and analyzing website page views.

Through this paper we can conclude that for the arbitrary given website we need to first collect the visitor page views history either through the website log data or through the pre-installed website analytics software like Google Analytics, then on the collected time series data we need to apply the smoothing technique to remove the irregularities and seasonal trend, And then we need to apply the regression technique to generate the future trends. The complexity is definitely is with the age of the website, as the newly developed website will not have the enough time series data to analysis. In such cases the comparison of similar website traffic may help for prediction.

The algorithm that needs to be developed through this methodology will be very complicated and also has lot of arithmetic's calculation. We will continue to study more on this area and tries to develop the rigid algorithm that can be easily applied to any website.

REFERENCES

[1] Chandana Napagoda - "Web Site Visit Forecasting Using Data Mining Techniques" International Journal Of Scientific & Technology Research Volume 2, Issue 12, December 2013. (http://www.ijstr.org/final-print/dec2013/Web-Site-Visit-Forecasting-Using-Data-Mining-Techniques.pdf) [published]

[2] Dan Peskin - "Back to the Future: Forecasting Your Organic Traffic" Moz.com [Blog] March 2013. (https://moz.com/blog/back-to-the-future-forecasting-your-organic-traffic)

[3] Tech-MBA - "Forecasting traffic for a website" in Tech-mba.com [Blog]. http://www.tech-mba.com/forecasting-traffic-for-a-website.html

[4] D. Ciobanu, C. E. Dinuca - "Predicting the next page that will be visited by a web surfer using Page Rank algorithm," in International Journal of Computers and Communications, 2012, pp.60-67

[5] T. Wang and Y. Ren, "Research on personalized recommendation based on web usage mining using collaborative filtering technique," WSEAS Trans. Info. Sci. and App., vol. 6, no. 1, pp. 62–72, Jan. 2009.

[6] W. Tong and H. Pi-lian, Web Log Mining by an Improved AprioriAll Algorithm, ;in Proc. WEC (2), 2005, pp.97-100.

[7] X. Wang, A. Abraham, and K. A. Smith, "Intelligent web traffic mining and analysis," J. Netw. Comput. Appl., vol. 28, no. 2, pp. 147–165, Apr. 2005.

[8] Professor Hossein Arsham "Regression Analysis with Diagnostic Tools for Predictions" ubalt.edu. (https://home.ubalt.edu/ntsbarsh/business-stat/otherapplets/Regression.htm)

[9] K. Driessens, "GaussianProcesses." Sourceforge.net . (http://weka.sourceforge.net/doc.dev/index.html?weka/classifiers/functions/GaussianProcesses.html) [Accessed: 6-Aug-2012].

[10] E. Frank and L. Trigg, "LinearRegression." Sourceforge.net. (http://weka.sourceforge.net/doc/weka/classifiers/functions/LinearRegression.html) [Accessed: 10-Aug-2012].

[11] M. Ware, "MultilayerPerceptron." Sourceforge.net. (http://weka.sourceforge.net/doc/weka/classifiers/functions/MultilayerPerceptron.html). [Accessed: 8-Aug-2012].

[12] E. Frank, L. Shane, and S. Inglis, "SMO." Sourceforge.net. (http://weka.sourceforge.net/doc/weka/classifiers/functions/SMO.html) [Accessed: 7-Aug-2012].

[13] M. Hall, "Time Series Analysis and Forecasting with Weka - Pentaho Data Mining." Pentaho.com. (http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka) [Accessed: 10-Aug-2012].

[14] Wikipedia "Linear regression" in Wikipedia.org. (https://en.wikipedia.org/wiki/Linear_regression)

[15] Wikipedia "Multilayer perceptron" in wikipedia.org. (https://en.wikipedia.org/wiki/Multilayer_perceptron)

[16] Jia Li and Andrew W. Moore "Forecasting Web Page Views: Methods and Observations" in Journal of Machine Learning Research 9 (2008) 2217-2250 October 2008.

[17] Z. Markov, D. T. Larose, Data Mining The Web Uncovering Patterns in Web Content, Structure and Usage. USA: John Wiley & Sons, 2007.

[18] H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, 3rd ed. Morgan Kaufmann, 2011.

[19] Improbable, Random House, New York, NY

[20] Sagar S. Badhiye, Kalyani S. Hatwar, P. N. Chatur - "Trend based Approach for Time Series Representation" International Journal Of Computer Applications (0975 – 8887) Volume 113 No. 16, March 2015. (http://research.ijcaonline.org/volume113/number16/pxc3901991.pdf) [published]

[21] Venkat Reddy - Data Analysis Course Time Series Analysis & Forecasting (Version-1) in slideshare.net [http://www.slideshare.net/21_venkat/timeseries-forecasting]