Web Site Visit Forecasting Using Data Mining Techniques

Chandana Napagoda

***Abstract*:** Data mining is a technique which is used for identifying relationships between various large amounts of data in many areas including scientific research, business planning, traffic analysis, clinical trial data mining etc. This research will be researching applicability of data mining techniques in web site visit prediction domain. Here we will be concentrating on time series regression techniques which will be used to analyse and forecast time dependent data points. Then how those techniques will be applied to forecast web site visits will be explained.

Keywords: Forecasting, Web Site, SMO Regression, Linear Regression, Gaussian Regression and Multilayer Perceptron

I. INTRODUCTION

In contemporary world with the technological changes, ability of predicting web site visiting patterns have a significance value for every site owner for targeting their business for right customers at right time. The available solutions to predict future access and usage patterns for web sites are Business Intelligence (BI) tools. But for many small and medium sized companies or web site owners are unable to afford for them since they are expensive solutions. The available BI tools including Google Analytic doesn't support forecasting features. They are focused on analysing user behaviours on the web sites and log related results. So the solution going to be developed from this research is targeted on web site owners and administrators to assist the future predictions of web site visits on their marketing strategies. 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 deallocates servers. All the above requirements exist in web site visit prediction domain motivated me on this research work for providing a suitable solution implementation. The aim of this research work is applying a suitable forecasting technique to predict web site visits. Thus, the results derived through forecasting will be assist web site owners in,

- Predict total number of visitors within next WEEK time.
- Predict number of visitors on a given DAY (Sunday, Monday, etc) within next WEEK time.

The objectives of this research work for achieving the aim are,

- Identifying and investigating data mining techniques which can be used for time series data forecasting.
- Identifying and applying suitable pre-processing techniques to clean the data.
- How the identified techniques can be applied in web site visit predication domain

- *Chandana Napagoda*
- *Department of Computer Science, University of Moratuwa Katubedda, Sri Lanka*
- *cnapagoda@gmail.com*

II. REVIEW OF LITERATURE

Forecasting is analysing and predicting the future behaviour of the selected data set. In Forecasting knowledge from the analysed data is used to predict future behaviours. It help in many ways in various domains such as controlling load balance, future marketing campaigns, allocating or de-allocating resources and caching ,prefetching web pages for improve performance. There are limited number of researches have been done on Web site related forecasting. D. Ciobanu, C. E. Dinuca done a research for Predicting the next page to be visited by a web user[1]. 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 web site log analysis to determine probabilities of visiting the pages. Their concept founded from the web page ranking algorithm Page Rank [2]. 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 [3]. 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 searching engines. Further they offer an improved algorithm based on the original AprioriAll algorithm, which has been used in Web, logs mining widely. Test results show the improved algorithm has a lower complexity of time and space [5]. 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 forecasting area [4]. The wide usage of the Internet in various fields has increased the automatic extraction of the log data from the web sites. The usage of data mining techniques on those data collected from the web helps us in pattern selection, which acts as a traditional way of decision-making tools. But with availability of current technologies we can use data from Application such as Google Analytics, and Alexa for pattern identification and forecasting.

III. PREPROCESSING

Data per-processing is an impotent step in data mining. It helps to remove garbage information effect from initial data set. It intends to reduce some noises, incomplete and inconsistent data. The results from pre-processing step can be later processed by data mining algorithms. In a data set of web site visits, marketing or related activities can create a sudden increase in number of visitors. But it won't retain for a longer time period. In this type of a situation that sudden increased visitor count data is considered as outliers.

A. Handling Outliers

The best way to remove outliers is organizing similar values in to group/ cluster. Outliers can be identified through human or computer inspection. But used human inspection to identify outliers patterns. For example consider a situation where mean page view count is 121, but some days has page view amount as 518. In this case, for removing the outliers fist calculate mean value and standard deviation (is the square root of variance). In a normal distribution curve, from μ–σ to μ+σ contains about 68% of the measurements. There for if the forecasting doesn't interested about outliers, the values which are out of the μ–σ to μ+σ range will be removed and assigned with mean value.

IV. FORECASTING TECHNIQUES

The data present in the environment of web site 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 forecasting'. Time series data such as web site 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. So one approach of handling time series data is removing its temporal ordering so that 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[11]. In this research we used minimum lag value as 1 and maximum lag value as 14.

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 [6]. Weka (>= 3.7.3) 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 trends and seasonality. After the data has been transformed, any of Weka's regression algorithms can be applied to learn a model. One approach is to apply multiple linear regressions. Also Weka allows any method capable of predicting a continuous target. For example support vector machine (SVM) for regression trees and model trees can be applied. Model trees are decision trees with linear regression at the leaves. Weka is consist with multiple classifier functions and in this research four of those classifier functions will be used. They are namely Gaussian Process, Multilayer Perceptron, Linear Regression and SMO Regression.

1. Gaussian Process

This classifier function implements Gaussian process for regression without hyper-parameter tuning. Here the missing values are replaced with global mean/mode values [7].

2. Multilayer Perceptron

This classifier function uses backpropagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time [8].

3. Linear Regression

This function uses linear regression for predication. Weighted instances are possible to be used with this approach [9].

4. 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 [10].

V. WEB SITE VISIT FORECASTING

This paper discusses about the Web site visitors forecasting for a web site owned by one of a Software Company. For this purpose we have used daily page views for the web site from the Google Analytics from 01 April 2011 to 22 July 2012 time range. Here a number of different prediction approaches have been applied such as Gaussian processes, linear regression, multilayer perceptron and SMO regression which are available on Weka time series analysis environment.

A. Date Set

We have consecutive 476 days of page view information which are downloaded from Google Analytic. The same page view data set cannot be used in prediction modelling and model validation both scenarios. Therefore information on last 14 days has been removed from the input data set and that data set is used to validate output of the prediction. The input data set is with only one attribute which is daily page view. Further the input data set will be used in forecasting in two different ways. Data set with outliers and data set without outliers are used in forecasting and variation of the forecast results will be evaluated along with the data set.

B. Forecasting Execution Environment

In this research, Waikato Environment for Knowledge Analysis-version 3.7.6 (Weka 3.7.6) tool and a package called "Time series forecasting environment" are being used for performing the predication modelling and predicting[11]. The input data set prepossessing and Attribute-Relation File Format (ARFF) generation steps are done manually and "Time series forecasting environment” is used for forecasting. Figure 1 shows structure of ARFF file.

```
@relation 'Analytics Visitors Overview'
@attribute Visits numeric
@attribute Day date 'MM/dd/yyyy'
@data
99,4/1/2011
67,4/2/2011
72,4/3/2011
102,4/4/2011
110,4/5/2011
```

Figure 1: View of Attribute-Relation File Format

C. Forecast Results

The result of the forecasted information is shown in the Table 1 and Table 2. Those results include values returned by applying each of the predication methods and actual page visit count. Hence result page visit information can be compared between each prediction methods; Gaussian, Linear regression, Multilayer Perceptron regression and SMO regression. As above mentioned this forecast has used page view information of 472 days.

TABLE I
FORECASTING RESULT WITH ACTUAL VALUES WITHOUT OUTLIERS

	SMO	Gaussian Regression	Linear Reg	Multilayer Perceptron	Actual Visits
06/30/12	82	82	82	72	94
07/01/12	77	77	84	61	59
07/02/12	119	113	123	108	126
07/03/12	119	117	121	105	125
07/04/12	118	118	120	113	132
07/05/12	118	117	122	110	109
07/06/12	113	110	115	100	126
07/07/12	79	76	79	70	51
07/08/12	79	75	83	63	64
07/09/12	118	111	122	110	105
07/10/12	118	116	120	111	146
07/11/12	117	115	119	111	121
07/12/12	118	115	121	112	117
07/13/12	113	108	114	88	108

TABLE II
FORECASTING RESULT WITH ACTUAL VALUES WITH OUTLIERS

	SMO	Gaussian Regression	Linear Reg	Multilayer Perceptron	Actual Visits
06/30/12	80	67	88	47	94
07/01/12	84	74	76	542	59
07/02/12	119	107	111	208	126
07/03/12	115	131	143	595	125
07/04/12	144	119	136	1086	132
07/05/12	128	150	152	1072	109
07/06/12	116	121	139	834	126
07/07/12	77	70	95	570	51
07/08/12	80	70	79	355	64
07/09/12	119	110	113	-64	105
07/10/12	122	129	145	-343	146
07/11/12	123	119	137	22	121
07/12/12	125	141	154	571	117
07/13/12	116	118	141	465	108

D. Results Evaluation

TABLE III
PERFORMANCE COMPARISON OF FORECASTING RESULT WITHOUT OUTLIERS

	MAE	MSE	RMSE	RBE	RSE
SMO Regression	0.357	3003	54.779	-0.0019	-14.577
Gaussian Processes	-2.237	2939	54.212	0.0096	-12.045
Linear Regression	3	3394	58.258	-0.01	-20.082
Multilayer Perceptron	-10.643	4387	66.234	0.0295	-12.186

TABLE IV
PERFORMANCE COMPARISON OF FORECASTING RESULT WITH OUTLIERS

	MAE	MSE	RMSE	RBE	RSE
SMO Regression	4.643	3411	58.403	0.0318	-23.36
Gaussian Processes	3.071	4617	67.9482	-0.0182	-27.48
Linear Regression	16.143	7848	88.589	1.0761	523.2
Multilayer Perceptron	319.786	3766961	1940.866	0.074	883.01

The accuracy of the above forecast results has been analysed and evaluated based on Mean Squared Error (MSE), Mean Absolute Error (MAE), Rooted Mean Squared Error (RMSE), Relative Mean Squared Error (RMSE) and Relative Absolute Error (RAE). Table 3 and Table 4 are shown with results of accuracy for each of the four forecasting methods used. Actual visit and forecasted graph representation shown in Fig. 1.

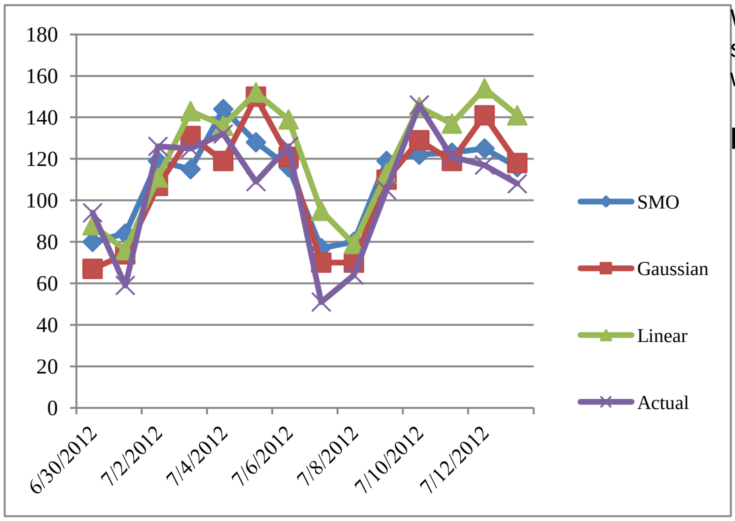

Fig. 1 A line graph representation of forecast result wit actual visit

The mean squared error and the mean absolute error shows how forecast result and actual output are closer. The relative mean squared error and relative absolute error are used as a good measure of accuracy. By comparing the results which are shown in Table 3 and Table 4, SMO Regression and Linear Regression are the best suited methods for predicting or forecasting Web site visit related information. When using Linear Regression, performing outlier's analysis is a very important pre-processing step which produces more accurate results. By comparing the results which are shown in Table 3 and Table 4, SMO Regression and Linear Regression are the best suited methods for predicting or forecasting Web site visit related information. When using Linear Regression, performing outlier's analysis is a very important pre-processing step which produces more accurate results.

VI.CONCLUSIONS

This paper had discussed about the Web site visitors forecasting for a web site owned by one of a Software Company. The data present in this environment is time dependent series of data points. Modeling and forecasting in this environment is Time Series analysis and Time Series Forecasting. So the paper had discussed about applicability of Weka tool which provides a knowledge analysis environment. Basically the research went along four regression algorithms which are applied against the data set with outliers and without outliers. So how the data pre-processing, particular algorithm will assist in achieving accurate results were evaluated end of the paper based on the derived results. Based on the evaluation results we can conclude that SMO regression and Linear Regression algorithms are better suited for forecasting web site related information and also using Linear Regression on pre-processed data gives more accurate results. Further when forecasting is done on web site related information, we can reduce various web hosting and server maintenance cost through prior acknowledging about future events. This research can be extended further with applicability of forecasting and data smoothing technology like Exponential smoothing.

ACKNOWLEDGMENT

We wish to acknowledge Mark Hall for developing Weka "Time series forecasting environment" package and given me advice we conduct this research.

REFERENCES

[1] 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*

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

[3] 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.

[4] 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.

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

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

[7] K. Driessens, “GaussianProcesses.” [Online]. Available: http://weka.sourceforge.net/doc.dev/index.html?weka/classifiers/functions/GaussianProcesses.html. [Accessed: 6-Aug-2012].

[8] M. Ware, “MultilayerPerceptron.” [Online]. Available: http://weka.sourceforge.net/doc/weka/classifiers/functions/MultilayerPerceptron.html. [Accessed: 8-Aug-2012].

[9] E. Frank and L. Trigg, “LinearRegression.” [Online]. Available: http://weka.sourceforge.net/doc/weka/classifiers/functions/LinearRegression.html. [Accessed: 10-Aug-2012].

[10] E. Frank, L. Shane, and S. Inglis, “SMO.” [Online]. Available: http://weka.sourceforge.net/doc/weka/classifiers/functions/SMO.html. [Accessed: 7-Aug-2012].

[11] M. Hall, “Time Series Analysis and Forecasting with Weka - Pentaho Data Mining.” [Online]. Available: http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka. [Accessed: 10-Aug-2012].