Time Series modeling of visitors' type on web analytics

Mohammad Amin Omidvar[1], **Vahid Reza Mirabi**[2], **and Narjes Shokry**[3]
[1]Information Technology Management, IAU E-Campus, Tehran, Iran
[2]Faculty of Management, IAU Central Branch, Tehran, Iran
[3]Faculty of Public Administration, IAU Central Branch, Tehran, Iran

Abstract - *The aim of this paper is to develop a flexible methodology to analyze the effectiveness of different variables on various dependent variables which all are times series specifically visitors' type on page views. This survey shows how to use a time series regression on one of the most important and primary index (page views per visit) on Google analytic and in conjunction it shows how to use the most suitable data to gain a more accurate result. There are too many data available on web analytics which are overwhelming for data analyzer. With this method, more accurate results are available. This methodology is critical for effective website monitoring and benchmarking that may lead to better website strategies. The value of this paper relies on introducing and using a systematic flexible methodology to analyze visitors' behavior and their impact on page views. Additionally this methodology can be used to analyze other time series variable.*

Keywords: worldwide web, Systems analysis, Data mining, visitors' behavior, web analysis, web metric, Google Analytics

1 Introduction

The internet is growingly rapidly and has a great impact on many businesses. Thousands of companies now own a website and websites have become an integrated part of the business. Furthermore many companies have employed many technologies which are available through the web such as online services. With web information, web developers and designers can improve user interfaces, search engines, navigation features, online help and information architecture and have happier visitors/ costumers [14]. One of the most popular ways which most frequented websites use to collect data and information about their websites is through web analytic. Web analytic collects a large amount of data from users such as browser type, connection speed, screen size, visitors' type, and etc. The collected data are usually large in quantity and type that need to be further processed to become useful information or knowledge.

1.1 Profile of the website

In 1998 an Iranian visual artist website was launched (http://www.omidvar.net). This website has many pages with images and few texts. The Google Analytics traffic overview showed that all traffic sources sent a total of 27,422 visits from 1 June 2008 to 31 March 2011. The total page views during this period were 145,874.

2 Impact of the internet on business

The internet has been playing the important role of corporate marketing during the past ten years [30]. With its combination of rich text, multimedia and user involvement, the internet contains more information than any other media [18], [25]. The internet offers speed, reach, and multimedia advantages, and has changed the way in which businesses interact with their customers, suppliers, competitors, and employees [8].

Nearly all businesses now have a website [12]. A corporate website enriches the image of a business and provides direct benefits in terms of electronic commerce (e-commerce) sales [22] and indirect benefits in terms of information retrieval, branding, and services [23]. Recognition of the internet is driving marketers in traditional companies to conduct transactions on the internet [10]. Barua, Konana, B.Whinston, & yin (2001) found that e-business operational excellence results in financial performance [5]. Thousands of companies have a fear to be left behind by their competitors if they do not use online technologies.

The total number of internet users and the number of websites are increasing significantly which will result in the rapid growth of the use of the World Wide Web for commercial purposes[16] [11] [9].

On a five year forecast by Forrester, e-commerce sales in the U.S. will grow 10% annually from 2014 and online retail sales will be nearly $250 billion, up from $155 billion in 2009 [27].

The increasing amount of internet users, websites and retail sales implies that web developing should be carried out in a competent, professional manner to increase profit.

However, systematic analysis of costumers' behaviors has not kept pace with the rapid growth in e-commerce. Without quantifiable metrics which is available through web analytics software, website optimization (WSO) is a guessing game, therefore a majority of e-commerce companies cannot afford this risk given their huge amount of money. Above 70% of the most frequented websites use web analytic tools but with their large amount of data, it is difficult to use them

effectively [1]. Therefore, it is important to understand what kind of data and knowledge are required for successful website development work.

Web Analytics Association Standards (2006) committee defined the three most important metrics as Unique Visitors, Visits/Sessions, and Page Views; and, also categorized search engine marketing metrics through counts (visits…), ratios (page views per visits….), and key performance indicators (KPIs) [3].

The main reasons for measuring Search engine marketing (SEM) successes are related to traffic measurement and the return on investment come on 4th [28]. The top for reasons are as follows:

Increased traffic volume (76%)
Conversion rates (76%)
Click-through rates or CTRs (70%)
Return on investment (67%)

Progressive improvement of SEM campaigns, conversion rates, and website performance are available through web metrics, which would results in an increase in profits, happier customers, and higher return on investment (ROI) by tracking progress over time or against the competition [4].

Online technology collects large amounts of detailed data on visitor traffic and activities on websites, which would cause a plethora of metrics [13], and on the other hand this variety of measures can be overwhelming. Developing a website is a dynamic ongoing process which is guided by knowledge of its visitors.

3 Methodology

Google Analytics allows users to export report data in Microsoft Excel format, which when transformed can be analyzed with time series statistical programs. The software EViews is used to compute time series regression [7]. Initially a data set with 27,422 entries for 34 months drawn from Google Analytics was employed to analyze the performance of page views or page views per visits which is defined as one of the three most important metric [2]. Monthly data series was the most suitable series among daily and weekly because the accuracy and credibility of the regression was higher than those of other series [6] [26].

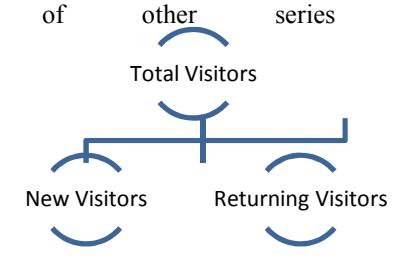

Figure 1 Independent Variables

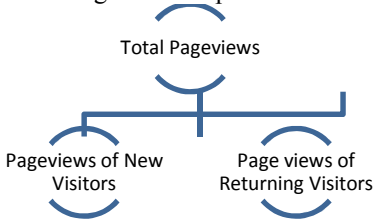

Figure 2 Dependent Variables

Before creating the model, some statistical matters with regards to the use of Google Analytics data in combination with time series methodology must be considered. For this reason all independent variables were processed by Augmented Dickey-Fuller Test to see if they were stationary or not. If the variables had a unit root they would be transformed to stationary by Difference-Stationary Process (DSP) [21].

Table 1 Augmented Dickey-Fuller test statistic (Level 10%)

Independent Variable	Has a Unit Root	Stationary After DSP
Total Visitors	No	
New Visitors	No	
Returning Visitors	No	

The first model is created with Autoregressive Moving Average (ARMA) model which consider total visitors as independent variable and total page views as dependent variable.

Table 2 Regression of Total Visitors

Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11 Time: 11:40
Sample: 2008M06 2011M03
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TOTAL_VISIT	3.125487	0.383079	8.158848	0.0000
C	1769.614	354.1795	4.996377	0.0000

R-squared	0.675347	Mean dependent var	4290.412
Adjusted R-squared	0.665202	S.D. dependent var	1744.983
S.E. of regression	1009.678	Akaike info criterion	16.72967
Sum squared resid	32622384	Schwarz criterion	16.81946
Log likelihood	-282.4044	Hannan-Quinn criter.	16.76029
F-statistic	66.56681	Durbin-Watson stat	1.662628
Prob(F-statistic)	0.000000		

Table 2 represents the following model and the coefficient of total visits can be considered as the average impact of visitors (3.13).

$$PAGEVIEWS = 3.13*Total\ VISITS + 1769.61 \quad (1)$$

The new model credibility and reliability is checked following these seven steps [19] [20] [15] [24].

1. Regression line must be fitted to data strongly ($R^2>0.6$).
2. Independent variables should be jointly significant to influence or explain the dependent variable (i.e. F-test, Anova)
3. Most of the independent variables should be individually significant to explain dependent variable (i.e. T-test).

4. The sign of the coefficients should follow economic theory or expectation or experiences or intuition.
5. No serial or auto-correlation in the residual (Breusch-Godfrey serial correlation LM test : BG test)
6. The variance of the residual (u) should be constant meaning that homoscedasticity (Breusch-Pegan-Godfrey Test)
7. The residual (u) should be normally distributed (Jarque Bera statistics).

Table 3 credibility and reliability of total visitors' regression

Independent Variable	Total Visitors
Dependent variable	Total Page views
R^2>0.6	Yes
Independent variables are jointly significant	Yes
Independent Variables are individually significant	Yes
The sign of the coefficients follow economic theory	Yes
No serial in the residual	Yes
The variance of u is constant	No
Normal distribution of u	Yes

For further information about the impact of other visitors such as new visitors and returning visitors on page views, we would normally consider total page views as dependent variable and the two mentioned variables as independent variables.

Table 4 Ordinary Regression for total page views

Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11 Time: 11:48
Sample: 2008M06 2011M03
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
NEW_VISIT	2.700648	0.420936	6.415813	0.0000
RETURNING_VISIT	7.132673	2.002317	3.562209	0.0012
C	1469.094	368.8106	3.983330	0.0004

R-squared	0.713622	Mean dependent var	4290.412
Adjusted R-squared	0.695146	S.D. dependent var	1744.983
S.E. of regression	963.4680	Akaike info criterion	16.66305
Sum squared resid	28776387	Schwarz criterion	16.79773
Log likelihood	-280.2719	Hannan-Quinn criter.	16.70898
F-statistic	38.62424	Durbin-Watson stat	1.489005
Prob(F-statistic)	0.000000		

Table 5 Credibility and reliability of ordinary total visitors' regression model

Independent Variable	Visitors' type (new visitors and returning visitors)
Dependent variable	Total Page views
R^2>0.6	Yes
Independent variables are jointly significant	Yes
Model	2.70*New visitors + 7.13*Returning Visitors + 1469.09

However, this method is not logical for web analytic because the detailed information of each visitor's page views is available [26]. Therefore total page view is broken down to its elements (page views of new visitors and returning visitors). Then, Autoregressive Moving Average (ARMA) model were created with their independent variables and their dependent variables. Since these models were describing part of total page views, only there fitness to data and significance were checked.

Ultimately, a new ARMA model of total page views was created with the sum of its related fundamental model. With the combined use of both dependent variables and independent variables, more accurate results were achieved.

Table 6 Regression model for each independent variable

Independent Variable	New visitors	Returning Visitors
Dependent variable	Page views of Search visitors	Page views of Direct visitors
R^2>=0.6	yes	Yes
Independent variables are jointly significant	Yes	Yes
Model	2.63* New visitors + 1458.77	5.32* Returning Visitors + 321.1

Table 7 new Regression for total page views

Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11 Time: 12:25
Sample: 2008M06 2011M03
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
2.63*NEW_VISIT	1.026862	0.160052	6.415813	0.0000
5.32*RETURNING_VISIT	1.340728	0.376375	3.562209	0.0012
C	1469.094	368.8106	3.983330	0.0004

R-squared	0.713622	Mean dependent var	4290.412
Adjusted R-squared	0.695146	S.D. dependent var	1744.983
S.E. of regression	963.4680	Akaike info criterion	16.66305
Sum squared resid	28776387	Schwarz criterion	16.79773
Log likelihood	-280.2719	Hannan-Quinn criter.	16.70898
F-statistic	38.62424	Durbin-Watson stat	1.489005
Prob(F-statistic)	0.000000		

The new model for total page views is as follow:

(2)

Total page views = 2.63*New Visitors + 5.32*Returning Visitors + 1469.1

Table 8 Credibility and reliability of new total visitors' regression model

Independent Variable	Visitors' Type
Dependent variable	Total Page views[1]
R^2>=0.6	Yes
Independent variables are jointly significant	Yes
Independent Variables are individually significant	Yes
The sign of the coefficients follow economic theory	Yes
No serial in the residual	Yes
The variance of u is constant	Yes
Normal distribution of u	Yes

4 Results and Analysis

What can be manifest from close analysis of the information in the Table 4 and Table 7 is that there is significant difference on returning visitors' impacts and their method. Initially, the impact of returning visitors was 7.13 and on new model it decreased to 5.32. The latest model is more appreciated because it measures the correct impact of independent variables on dependent variables.

Many websites have measured their success with their competitors' behavior and many others have measured that with their own website success. This survey had introduced a methodology to measure the success of the website with its time series data. It also focused on one of the most primary and important variables which are page views and page views per visits, and showed how to use the most suitable data for measurment. This method can be used on all websites and time series variables.

5 Recommendation

One of the most difficult variables in webanalytic is returning visitors because it requires users to enable cookie on their browser, not delete the past cookie files, and also avoid changing browser. Some sites has used Ip address, username, and cookies all together to overcome this problem, but there are still problems in considering them as returning visitors. For Example, if a new visitor use some one else computer which has not logged out from the site, with all efforts that is done to categorise him correctly on returning visitors, ironicaly he will be labeled as returning visitors because he has used the same user name, IP address and cookie. However, it would make sense to consider direct visitors as returning visitors, since they have some simmilarities in behavior [26] .

This methodology can be used on more detailed variables, such as new visitos from refferal sites with T1 speed to get more detail information about users behaviour. Furthermore, since Google Analytics is integrated with Adwords and Adsense and their time series data's are available on Google Analytics, so further studies on time series data of Adsense Revenue or Adwords campaigns might have interesting results; because they bring traffic and revenue to the site.

Although this methodology is not recommended for search visitors' behavior and some semantic term is suggested [17], combining some other attributes such as visitors speed, terriority and type might help to perdict search visitors behaviour even without knowing the semantic terms of quarries.

6 Discussion

Many website owners or developers have used web traffic elements for their website performance, but the important thing is the knowledge gained about visitors and their behavior to keep them happy and satisfied with the website. Most of the websites have used web analytics to collect data about visitors with no systematic way to convert these data into tangible knowledge. The amount of these data which is available through web analytic is immense, and the developer may get lost in it. So a systematic way is needed to analyze these data.

7 References

[1] Google Analytics Usage Statistics. (2010, Jun). Retrieved from Web and Internet Technology Usage Statistics: http://trends.builtwith.com/analytics/Google-Analytics

[2] Association, W. A. (n.d.). "Big Three Definitions" Ver. 1.0. 2300 M Street, Suite 800, Washington DC 20037, United States.

[3] Association, W. A. (2006). Web Analytics "Big Three" Definitions., (p. 2). Washington DC 20037.

[4] B.king, A. (2008). Website Optimization. Orielly.

[5] Barua, A., Konana, P., B.Whinston, A., & yin, F. (2001.). Driving e-business excellence. Sloan Management, Rev. 34(1) 36–44.

[6] Batchelor, P. R. (n.d.). EVIEWS tutorial:Cointegration and error correction. City University Business School, London.

[7] Beatriz Plaza Faculty of Economics, U. o. (2009). Monitoring web traffic source effectiveness with Google Analytics An experiment with time series. Emerald, 9.

[8] Bodily, S., & Venkataraman, S. (2004). Not walls, windows: capturing value in the digital age. Journal of Business Strategy, Vol. 25 No. 3, pp.15-25.

[9] capital, I. u. (n.d.). ITU World Telecommunication/ICT indicators database.

[10] Chakraborty, G. L. (2002). An empirical investigation of antecedents of B2B Websites' effectiveness. Journal of Interactive Marketing, 16: 51–72. doi: 10.1002/dir.10044.

[1] Total page views= page views of New Visitors + page views of Returning Visitors

[11] Consotium, I. S. (2010). The ISC Domain Survey | Internet Systems Consortium. Retrieved from Internet Systems Consortium | Internet Systems Consortium: http://www.isc.org/solutions/survey

[12] Cotter, S. (1993). TAKING THE MEASURE OF E-MARKETING SUCCESS. Journal of Business Strategy, Vol. 23 Iss: 2, pp.30 - 37.

[13] FMI Group. (2001). Web site Visitors Analysis-Statistics or Intelligence? Basinstoke: FMI Group.

[14] Fourie, I., & Bothma, T. (2007). Information seeking: an overview of web tracking and the criteria for tracking software. Aslib Proceedings, ISSN: 0001-253X, 24.

[15] Garson, G. D. (2010, 2009, 2008). Logistic Regression. Retrieved July 2010, from http://faculty.chass.ncsu.edu: http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

[16] Global Number of Internet Users,total and per 100 inhabitants 2000-2010. (n.d.). Retrieved from ITU: Committed to connecting the world: http://www.itu.int/ITU-D/ict/statistics/material/excel/2010/Internet_users_00-10.xls

[17] Gupta, Siddharth; Thakur, Narina;. (2010). Semantic Query Optimisation with Ontology. International journal of Web & Semantic Technology (IJWesT).

[18] Hoffman, D. L., & Novak, T. P. (October 1996). Marketing in Hypermedia Computer-Mediated Environments. journal of marketing, 50-68.

[19] Hossain, S. (2006, 6 16). An Investigation into Regression Model using EVIEWS. Retrieved from http://www.sayedhossain.com/: www.sayedhossain.com/files/Lec1.Regression.ppt

[20] Hossain, S. (n.d.). An Investigation into Regression Model using EVIEWS. Lecturer for Economics.

[21] http://www.hkbu.edu.hk/~billhung/econ3600/application/app01/app01.html. (n.d.). Dickey-Fuller Unit Root Test. Retrieved from Hong Kong baptist university: http://www.hkbu.edu.hk/~billhung/econ3600/application/app01/app01.html

[22] Inge Geyskens, K. G. (2002). The Market Valuation of Internet Channel Additions. The Journal of Marketing, Vol. 66, No. 2 (Apr., 2002), pp. 102-119.

[23] Lederer , A., Mirchandani, D., & Sims, K. (2001). The Search for Strategic Advantage from the World Wide Web. International Journal of Electronic Commerce, Volume 5 , Issue 4 Pages: 117-133 .

[24] Ludlow, E. (n.d.). Multiple Regression: Fitting Models for Multiple Independent Variables.

[25] Okazaki, S., & Rivas, J. A. (2002). A content analysis of multinationals' Web communication strategies: cross-cultural research framework and pre-testing. Internet Research, Vol. 12 No. 5, pp.380-90.

[26] Omidvar, Mohammad Amin; Mirabi, Vahid Reza; Shokri, Narjes. (2011). Analyzing the impact of visitors on page views with Google Analytic. International Journal of Web & Semantic Technology (IJWesT).

[27] Schonfeld, E. (2010, March 8). Forrester Forecast: Online Retail Sales Will Grow To $250 Billion By 2014. Retrieved from TechCrunch: http://techcrunch.com/2010/03/08/forrester-forecast-online-retail-sales-will-grow-to-250-billion-by-2014/

[28] Search Engine Marketing Professional Organization, S. (2008, February). Search engine marketing 2007.

[29] SEMPO. (2006). Search Engine Marketing Professional Organization survey of SEM agencies and advertisers.

[30] Welling, R., & White Macquarie, L. (2006). Web site performance measurement: promise and reality. Managing Service Quality, SSN: 0960-4529.