Decision Support Based on Time-Series Analytics: A Cluster Methodology

Wanli Xing[1], Rui Guo[2], Nathan Lowrance[1], and Thomas Kochtanek[1]

[1] School of Information Science and Learning Technologies, University of Missouri, Columbia, MO 65211, USA
{wxdg5,njl352}@mail.missouri.edu, KochtanekT@missouri.edu
[2] Department of Civil and Environmental Engineering, University of South Florida, Tampa, Columbia, FL 33620, USA
rui@mail.usf.edu

Abstract. Web analytic techniques have become increasingly popular, particularly Google Analytics time-series dashboards. But interpretations of a website's visits traffic data may be oversimplified and limited by Google Analytics existing functionalities. This means website mangers have to make estimations rather than mathematically informed decisions. In order to gain a more precise view of longitudinal website visits traffic data, the researchers mathematically transformed the existing Goggle Analytics' log data allowing the vectors of website visits per each year to be considered simultaneously. The methodology groups the data of an example website gathered over an 'x' year period into 'y' clusters of data. The results show that the transformed data is richer, more accurate and informative, potentially allowing website managers to make more informed decisions concerning promoting, developing, and maintaining their websites rather than relying on estimations.

Keywords: Temporal analytics, Google analytics, cluster analysis, decision support, website management.

1 Introduction

Web analytics are equally valuable for profit and nonprofit and many website mangers have turned to web analytics techniques to help them make more informed decisions about adverting, site development, and site maintenance [1]. Google Analytics (GA) has become a leading tool in this context and can provide quick access to metrics to ascertain traffic levels and visitor distribution [2]. There are some limitations to these metrics. One is Spider visits are indistinguishable from true visitors. Spiders are computer programs that access sites to update databases [3]. Also visits alone can lead to an overestimation of a site's visitor traffic because some visitors will leave if a page is having trouble loading or stay for too brief an amount of time to matter [4,5]. Second, when analyzing web metrics, it is important to remember that no firm inferences regarding user's intentions can be made solely from web metrics [6, 7]. These limitations aside, GA temporal metrics can still be informative.

S. Yamamoto (Ed.): HIMI 2014, Part II, LNCS 8522, pp. 217–225, 2014.

Temporal fluctuations of when visits are occurring have had noticeable effects on the interpretation of web traffic [8]. Categories of certain queries trend differently over varying periods of time, supporting the importance of temporal analysis [9]. Temporal factors also relate to the quality of web searches, name search effectiveness and efficiency [10]. Temporal analysis has also been applied to study the dynamics of blogger's posting behaviors [11]. Search engine transactional logs and time series analysis has been established as a viable means of anticipating future web traffic on sites [12]. Because of these benefits the GA time series dashboard is a frequently used tool to provide a rough estimate of overall trends of visits to web sites [13-16].

Current time series analysis of GA data is following two thematic paths. One of these is based on website managers observing the GA time traffic dashboard and using the visits chart to roughly estimate the overall trend of the visits to the website. These estimations do have their uses. In one study a website experienced a decline in usage and these GA visits data were used to help interpret the reasons behind the decline [16]. Another study made uses of two years' worth of GA visits data for a health professional education website to inform findings, allowing a trend to predict that this particular site would further expand to be a global source on genetics-genomics education [14]. The drawback is that these conclusions about their websites evaluated over time depended on observation and estimation of the GA dashboard rather than accurate computation. Because of this, interpretations of visitor's traffic data may be oversimplified [17] and subject to limitations endemic to Google Analytics' existing functionalities

The second theme of use for time series analysis of GA data is based on using regression analysis of the website visits traffic data over a certain period of time and comparing its relationship with other website metrics. Plaza [18, 19] tested the relationship examining the effectiveness of entries (visit behavior and length of sessions) depending on their traffic source: direct visit, in-link entries over an approximately two-year period. Wang et al. [20] studied over a one-year period whether users behave differently during weekdays and weekends, finding a number of significant relationships between several key traffic variables and web metrics.

Many website managers are not likely to run regression analysis on their page traffic, so a more precise quantitative method could be helpful in comparison with the current estimations. None-the-less website users' behaviors are very important information to understand the market demand and to make strategic plans for a web system. Their visiting patterns can vary significantly over a long period of time. In order for the web managers to make decisions effectively, marketing and maintenance plans should be created in suitable intervals that are in line with the visiting patterns. For example, the webmaster should promote any products or services when they have the highest visits to their website, but schedule maintenance for the website when there are the fewest visitors. How can we more precisely identify users' visiting patterns over a longitudinal period rather than using a rough estimation through GA Dashboard?

To answer this question, we investigated a data mining method to provide a longitudinal and accurate view for web managers to use so that they can make decisions more effectively. By capturing temporal features of a website users' behavior with a

mathematical method and analyzing each year simultaneously, clustered results should provide a more accurate grouping of high, low, and median traffic levels and their corresponding dates. This study utilizes the Truman library website (http://www.trumanlibrary.org/) to illustrate a proposed methodology, but the application of this methodology is not limited to this type of website, as any website evaluated with GA should be able to use this method. This approach transforms GA log data from a somewhat limited interpretive state to something that is richer, more accurate, and informative.

2 Methodology

2.1 Method

Cluster analysis is classified as data set into groups that are relatively homogeneous within themselves and heterogeneous between each other on the basis of a chosen set of variables [21]. Therefore, it served our purpose for identifying groups of time slots which the website master can depend on to make informed decisions for their web systems. A cluster methodology for pattern identification of the website is shown in the following figure (Fig 1). GA provides a rather solid basis to accomplish the cluster analysis because it automatically collects all the visits data and their associated visits time for us.

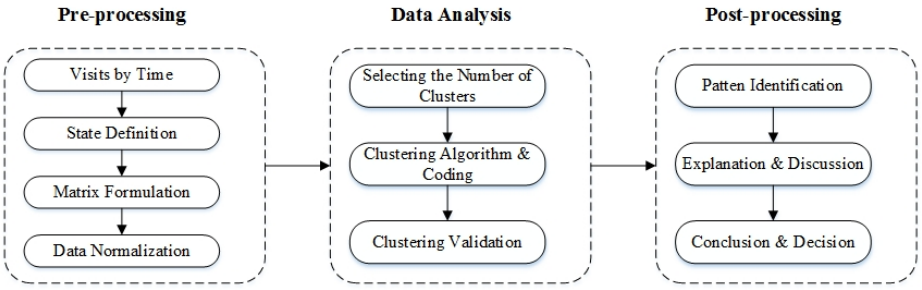

Fig. 1. Time-Series Analytics Framework

Pre-processing. In terms of gathering the visits by time, the first step is to access GA, which enables one to export data in different granularity such as hourly, daily, weekly, monthly and yearly based on the needs of the analysis.

The system state is an abstract representation of the condition of a system at some point in time [11]. Based on the different granularity of the data (e.g. week, month) collected by GA, we could capture the temporal feature of the website user's behavior with a mathematical method. The state definition used here is a vector of website visits in a certain time window (visits/day, visits/week etc.) and the time feature (a dimension of the time that website visits occurs) for each year. Since this research aims to look at analysis over a longitudinal period of time, for instance, identifying

user visiting patterns in one-year period. The data samples are thus multidimensional because the vectors of website for each year are considered simultaneously.

Therefore, the system states in our study are defined as follows, assuming there are M representative years in the log datasets and yearly visits are recorded into T time intervals. Then the data A, a K (M+1) × K (T) matrix, will have the format as following (1).

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} & a_{1(M+1)} \\ a_{21} & a_{22} & \cdots & a_{2M} & a_{2(M+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{(T-1)1} & a_{(T-1)2} & \cdots & a_{(T-1)M} & a_{(T-1)(M+1)} \\ a_{T1} & a_{T2} & \cdots & a_{TM} & a_{T(M+1)} \end{bmatrix} \quad (1)$$

In order to smooth the different data visits to scale among different years, the elements in this matrix should be properly normalized, which following equation is carried out prior to the cluster analysis so that make the data dimensionless (2).

$$atm' = atm - amsm \qquad (t=1,2,\cdots, T; \quad m=1,2,\cdots, M+1) \quad (2)$$

Where a_{tm}, $\bar{a}_m$ and s_m represents original, average, and standard deviation of website visits or time variables, respectively, for any particular observation.

Data Analysis. We choose K-means algorithm for the clustering analysis, one of the most popular non-hierarchical methods to do the analysis [25]. However, before moving on the specific K-means algorithm, one significant step is to decide the number of clusters we are going to use. The common practice before K-means clustering is to employ Gap-statistic to determine the proper number of clusters [22, 23]. The basic idea behind Gap-statistic is to find an "elbow" in the plot of the optimized cluster criterion against the number of clusters, K.

For this purpose, letting E_N^* denote the expectation under a sample size of N from the reference distribution, the optimal value for the number of clusters is then the value K for which the "Gap" is the largest. K is the number of clusters, N is sample size, and W_K denotes an overall average within the cluster sum-of-squares (3).

$$\mathrm{Gap}_N(\mathrm{K}) = E_N^* \{\log(W_K)\} - \log(W_K) \quad (3)$$

Those interested in the theoretical details of this method can refer to the Tibshirani, Walther, and Hastie's original paper [23]. In terms of the cluster algorithm details, in general, elements are grouped according to their similarities, and in K-Means cluster, the distance between them. In this study, squared Euclidean distance is employed to calculate the distance between clusters, where d_{ij}^2 is the squared Euclidean distance between state elements i, and j; x_{im} is the m^{th} element in state i; and $x_{jm}x_{ik}$ is the m^{th} element in state j (4).

$$d_{ij}^2=\sum_{m=1}^{M+1}(a_{im}-a_{jm})^2 \qquad (i,j=1,2,\cdots,T; \quad m=1,2,\cdots,M+1) \tag{4}$$

Vilifying the optimal number of clusters is also a critical step in cluster analysis. In our study, we implemented the Silhouette measure to vilify the efficiency of the selected number of clusters. For more Specific information on the Silhouette Coefficient refer to Rousseeuw's work [24].

Post Processing. After cluster analysis, post processing is conducted to determine the intervals and identify the visits' patterns. Another important step in our procedure is to explain the patterns. This requires the webmaster in conjunction with the context to infer meaning out of the pattern and make the informed decision.

2.2 Research Context

To demonstrate this approach, we chose a library website – Truman Presidential Library http://www.trumanlibrary.org/libhist.htm.

2.3 Dataset

Google Analytics was used to gather data over five years from August first, 2008 to July thirty-first, 2013. To better serve the purpose of the study, clickstream data pertaining to time was collected. We downloaded the CSV files containing weekly visits of website as shown bellow (Fig 2).

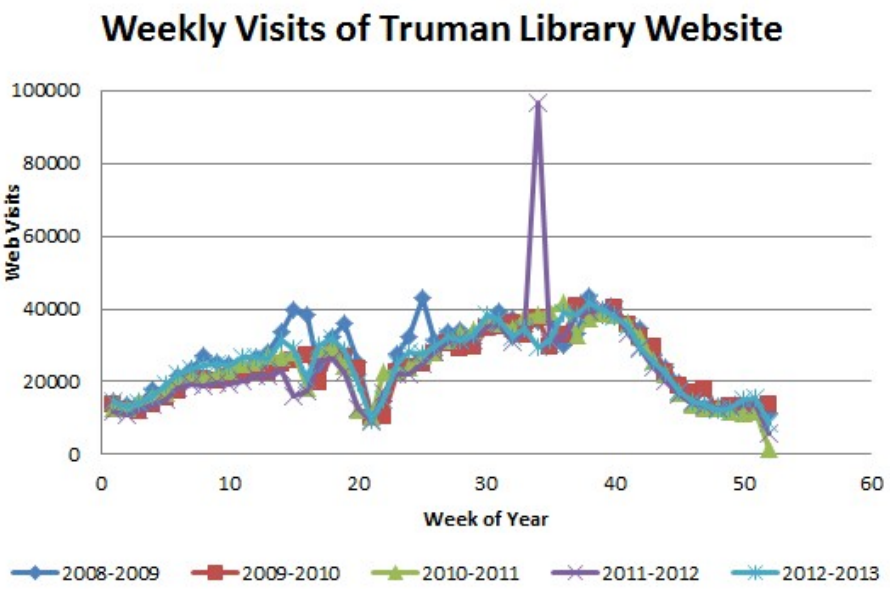

Fig. 2. Weekly Visits of Truman Library website

3 Results

3.1 Cluster Numbers

To choose the optimal number of clusters, the Gap statistic measure was conducted by coding in R. Below, we first show the observed and expected log (Wk) and compare the Gap values against the number of clusters in our case study (Fig. 3). Second, we show the Gap values against the number of clusters in our case study(Fig. 3). Due to these results we chose three as the number of clusters for K-Means algorithm in the next step.

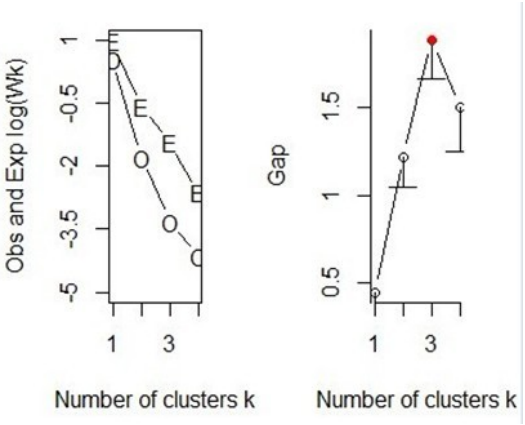

Fig. 3. Gap Function

3.2 Pattern Identification

The K-means clustering successfully identifies users' visiting patterns based on the average weekly visits and the time that activity is occurring (Table 1).

Table 1. Results of Cluster Analysis

Cluster (K=3)	Week
Cluster 1 (High)	34
Cluster 2 (Medium)	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 23, 24, 25, 26, 27, 28, 29 ,30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 32, 43, 44
Cluster 3 (Low)	1, 2, 3, 4, 5, 6, 7, 20, 21, 22, 45, 46, 47, 48, 49, 50, 51, 52

3.3 Cluster Validation

In order to validate the number of the selected clusters a Silhouette Coefficient was calculated. The "elbow" occurs when the number of clusters is three, which indicates the efficiency of our three cluster analysis.

3.4 Explanation

These results better inform the web manager of their peak traffic week, which would be week 34. Any important updating should be done before this high traffic week, it also informs the site manager of an optimal week for advertising or promoting if necessary. Cluster 3 indicates ideal times for site maintenance.

4 Discussion

Our methodology groups 52 weeks into different clusters based on the five years of click streaming data. These different clusters can help managers to develop plans for allocating resources in a more efficient fashion than making estimations based on looking at line graph data for individual years. For example, plans might be made for website maintenance, for adjusting time/task efforts and for the allocation of human resources better suited for expected website visiting conditions for particular dates or times. Therefore, managers, instead of basing decisions on a simple observation and estimation, obtain reliable quantified results that can be used to improve the quality of their decision-making. Keep in mind GA time series traffic dashboard only show a line graph of visit information chronologically (Fig. 4). While one could create an overlapping graph, like what was done for this paper (Fig. 2) and see major peeks and dips, it is still not precise. It is clear that some years have outliers, but our method clarifies grey areas, showing what are the true peaks across five years and what qualifies as the cut off points for high visit peaks and low visit dips. Ultimately this adds more mathematical confidence in the accuracy of high and low traffic times. This method offers a more precisely quantified and longitudinal point of view, and thus has the potential to be applied in different contexts and for different usages. They can simply redefine the abstraction state and time granularity. The next logical step would be to develop a tool to enhance GA allowing this mathematical clustering method to use GA visit data and be calculated for GA users, rather than hoping the site managers would be able and willing to use this procedure.

Fig. 4. GA view of visits to Truman Library website

5 Conclusion

As an exploratory study, this research presented an application of mathematical modeling using time series distributed click steaming data that GA collected for a library website to provide a more accurate account of past visits site traffic patterns. This new methodological framework based on advanced analytical techniques was developed to more accurately examine the visitors' behavior patterns over time, allowing for more confident web site management decisions to be made by site managers. In the future, we recommend the investigation of visitors' behavior patterns from different traffic sources (direct, reference or search engine) as well as return visitors' navigation in comparison with that of new visitors. This would help to identify the loyal users of a website, and other behavior characteristics could be further explored in an extended study as well.

References

1. Marek, K.: Getting to Know Web Analytics. Using web analytics in the library, pp. 11–16. ALA Store (2011)
2. Clifton, B.: Advanced web metrics with Google Analytics. John Wiley and Sons, Inc., Indianapolis (2012)
3. Phippen, A., Sheppard, L., Furnell, S.: A practical evaluation of Web analytics. Internet Research 14(4), 284–293 (2004)
4. Mullarkey, G.W.: Internet measurement data - practical and technical issues. Marketing Intelligence & Planning 22(1), 42–58 (2004)
5. Dreze, X., Zufryden, F.: Is Internet Advertising Ready for Prime Time? Journal of Advertising Research 38(3), 7–18 (1998)
6. Grimes, C., Tang, D., Russell, D.: Query Logs Are Not Enough. In: Workshop on Query Logs Analysis: Social and Technological Challenges, Banff, Canada (2007)
7. Weischedel, B., Huizingh, E.: Web Site Optimization With Web Metrics: A Case Study. In: International Conference on Electronic Commerce (ICEC 2006), pp. 463–470 (2006)
8. Khoo, M., Pagano, J., Washington, A.L., Recker, M., Palmer, B., Donahue, R.A.: Using web metrics to analyze digital libraries. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 375–384. ACM (2008)
9. Jansen, B.J., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management 42(1), 248–263 (2005)
10. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., Grossman, D.: Temporalanalysis of a very large topically categorized web query log. Journal of the American Society for Information Science and Technology 58(2), 166–178 (2007)
11. Zhang, Y., Jansen, B.J., Spink, A.: Time series analysis of a Web search engine transaction log. Information Processing & Management 45(2), 230–245 (2009)
12. Chi, Y., Zhu, S., Song, X., Tatemura, J., Tseng, B.L.: Structural and temporal analysisof the blogosphere through community factorization. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)
13. Farney, T., Mchale, N.: Data Viewing and Sharing: Utilizing Your Data to the Fullest. Library Technology Reports 49(4), 39–42 (2013)

14. Kirk, M., Morgan, R., Tonkin, E., McDonald, K., Skirton, H.: An objective approach to evaluating an internet-delivered genetics education resource developed for nurses: using Google Analytics™ to monitor global visitor engagement. Journal of Research in Nursing 17(6), 557–579 (2012)
15. Pakkala, H., Presser, K., Christensen, T.: Using Google Analytics to measure visitor statistics: The case of food composition websites. International Journal of Information Management 32(6), 504–512 (2012)
16. Kent, M.L., Carr, B.J., Husted, R.A., Pop, R.A.: Learning web analytics: A tool for strategic communication. Public Relations Review 37(5), 536–543 (2011)
17. Kumar, C., Norris, J.B., Sun, Y.: Location and time do matter: A long tail study of website requests. Decision Support Systems 47(4), 500–507 (2009)
18. Plaza, B.: Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. Aslib Proceedings 61(5), 474–482 (2009)
19. Plaza, B.: Google Analytics for measuring website performance. Tourism Management 32(3), 477–481 (2011)
20. Wang, X., Shen, D., Chen, H.L., Wedman, L.: Applying web analytics in a K-12 resource inventory. The Electronic Library 29(1), 20–35 (2011)
21. Guo, R., Zhang, Y.: Identifying Time-of-Day Breakpoints Based on Nonintrusive Data Collection Platforms. Journal of Intelligent Transportation Systems (2013)
22. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, 5th edn. John Wiley & Sons, Ltd. (2011)
23. Tibshirani, R., Walther, G., Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistic. Journal of Royal Statistical Society, B63, Part 2, 411–423 (2001)
24. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematic 20(1), 53–65 (1987)
25. Xu, C., Liu, P., Wang, W., Li, Z.: Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis & Prevention 47, 162–171 (2012)