Integration of the Google Analytics tool into the data pre- processing layer for WEB Usage Mining: A case study

Şükrü Can Şayan[1*], **Tahsin Çetinyokuş** [1]

1 Gazi University, * Corresponding author, tahsinc@gazi.edu.tr

Abstract

The web has to grow,interpret and analyze with each passing day.This has led to the emergence of the field of web usage mining. The methods, developed in this area, are trying to pre-process and analyze the server log files. Instead of this method, this study proposes Google Analytics integrated model for obtain data and develops an application that accelerates this process. The comparative results of the methods are included in the study.

Keywords: Web usage mining, Pre-processing, Google Analytics.

Citation: Şayan, Ş. C., Çetinyokuş, T. (2018, October) *Integration of the Google Analytics tool into the data pre- processing layer for WEB Usage Mining: A case study*. Paper presented at the Fifth International Management Information Systems Conference.

Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

IMISC 2018 PAPER

Integration of the Google Analytics tool into the data pre-processing layer for WEB Usage Mining: A case study

Abstract

The web has to grow, interpret and analyze with each passing day. This has led to the emergence of the field of web usage mining. The methods, developed in this area, are trying to pre-process and analyze the server log files. Instead of this method, this study proposes Google Analytics integrated model for obtain data and develops an application that accelerates this process. The comparative results of the methods are included in the study

Keywords

Web usage mining, Pre-processing, Google Analytics

1. Introduction

Today, web sites have become an indispensable tool for work and communication of any kind of organization. With web sites today, information can be collected or sent. Web technologies and conceptual progress have been experienced during last decade. Old web sites are predominantly text and partly visual, while today's media are predominant.

With the web technology that developed in recent years, the web has become the core of some processes and businesses. Like the developments in information and communication technologies, the increase in the amount of data collected and the development of analysis tools have increased the importance of firms to analyze the data they collect. Data mining applications seem to be spreading both in the academic field and in the business world (Seyrek & Ata, 2010).

According to the literature, there are many studies in this area. Also in the current literature there are many field type applications and written essays made in this area. Related studies have been searched and listed on Web of Science, IEEE Xplore and Google Scholar. The search terms "web usage mining" and various key words related to the field are used.

It’s found 767 results on 411, Web of Science in IEEE Xplore database. As a result of the investigations, the studies on IEEE Xplore have been observed to be more refined. When the studies are examined, it is seen that the data are generally collected and pre-processed. Some studies have focused specifically on preprocessing processes and have produced algorithms in this area.

2. Method

In general, it is observed that the server access logs are used in the studies. These records are extracted from the servers and analyzed and the improvement works are being done on the web sites. However, the static structure of these files, and their lack of detail, cause some problems to be encountered and require intensive pre-processing.

2.1. Problems with Access Logs Analyzing

Web servers adds log to file for every request they receive. The server records a log in the following format for each request as *remotehost rfc931 authuser [date] "request" status bytes* (World Wide Web Consortium, 1995)

Log files are kept as text and need to be parsed before processing. We are trying to develop solutions to these pre-processing problems that are experienced in this area. The main problems encountered in the analysis of logs are described in below.

User Identification: The analysis of Web usage does not require knowledge about a user's identity. However, it is necessary to distinguish among different users (Liu, 2007). The log files contain the IP address and the model name of the client. IP addresses, alone, are not generally sufficient for mapping log entries onto the set of unique visitors. This is mainly due to the proliferation of ISP proxy servers which assign rotating IP addresses to clients as they browse the Web (Liu, 2007).

Session Identification: Sites such as news sites can be visited once a day, and social media sites can be visited several times a day. Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site (Liu, 2007). To do this, all of the visit records are examined and the interactions are grouped by time values. Many commercial products use 30 minutes as a default timeout, and established a timeout of 25.5 minutes based on empirical data (Cooley, Mobasher, & Srivastava, 1999).

Data Cleaning: Search engines reads and records all pages of internet. Web usage mining is used to mine the user access behavior and for this reason, all log records relevant to non-human access have to be removed from the weblog (Sukumar, Robert, & Yuvaraj, 2016).

Data Integration: In addition to user and product data, e-commerce data includes various product-oriented events such as shopping cart changes, order and shipping information, impressions click-troughs, and other basic metrics primarily used for data analysis (Liu, 2007). Log files basically store page links and access, some criteria that could be of benefit in an analysis study cannot be recorded. For example, demographic information, address, and information that can benefit analysis such as sales amount, sales name, sales method, etc., should be merged later. However, the state of this records prevents them from being assigned to the right persons.

2.2. Google Analytics

Analytics, which was introduced 12 years ago, provides collection and analysis of web statistics. Analytics, it is also employed because Analytics is a free service offered by Google that generates detailed statistics about the visits to a website, and which is a user friendly application with the guarantee of Google technology.(Plaza, 2011). Analytics works with JavaScript code in client side.

2.2.1 Problems Solved Using Analytics

Analytics collects client-side information on each visit and sends them to their own servers, allowing them to analyze their data with their own web analytics tools or to create new analytics by experienced users. Another feature is API, which allows data to be retrieved from the system according to certain rules.

User Identification: Analytics uses cookies to identify users. Analytics JavaScript page tag, this information plus other visitor is collected and a set of cookies are created to identify the visitor (Clifton, 2012). With cookie, each user can be watched separately from each other. Another feature is that if the website can distinguish different users, this information can be used to combine different browsers that a person has used under a single user.

Session Identification: Analytics can distinguish itself by calculating sessions. The 30-minute rule is the unwritten standard across the web analytics industry (Clifton, 2012).

Data Cleaning: Analytics only saves real visits. Thus, the size of the stored data is as much as the required data size, and records without information are avoided.

Data Integration: The greatest strength of the Analytics product is the unification of different data. There are 699 properties on Analytics while the server registration files mainly hold 7 types of data. This data is collected under different categories and divided into metrics.

3. Study

The data pre-processing is the initial step in the data preparation process, aims to reformat the original logs to identify user's sessions. This process is most time consuming and intensive step (Reddy, Reddy, & Sitaramulu, 2013). This study, preferred the use of Analytics in the provision of data and was designed to integrate with Analytics, and to import data from the Analytics instead of server access logs. For this study, traffics of an e-commerce site between 01-23.07.2018 were used and the obtained data were tested with a mining modeling. Various measurements have been made for evaluate the performance of the proposed system and the data source. The website has been using Analytics since 2016 with the UserId feature. In order to make performance comparison, server access records were also calculated for the same period.

3.1. Google Analytics Integration as A Data Source

It is possible to setup Analytics integration and use existing integration for import access data. To achieve this, an application named GAEWKM has been developed. With this application, users can create their own data patterns and make modeling and analysis by obtaining necessary data from Analytics source. The application data selection interface is shown in Figure 1.

INSERT FIGURE 1 HERE

With the application developed, the analyst can use data from 699 different parameters by Analytics. With GAEWKM, users can obtain data without having to write a single line code. (Figure 2, Figure 3)

INSERT FIGURE 2 HERE

INSERT FIGURE 3 HERE

The developed application has been collected data for the mentioned interval. During this collection, the following 6 properties were determined.

- ga:dimension1: The custom dimension for user tagging and value changes between user-[userId] or visitor- [timestamp] values according to the state of the user.
- ga:sessions: Session count of user.
- ga:pageviews: Pageviews count of user.
- ga:avgSessionDuration: Average session duration of user.
- ga:transactions: The number of purchases of user.
- ga:transactionRevenue: Total trade amount in of user.

When these data were collected, 3 filters were selected.

- ga:dimension1: The value of the dimension should start with a user or visitor keyword.
- ga:sessions: Number of sessions must be greater than 3.
- ga:browser: Browser header is not null for eliminating bots.

Data selection after this process has been run 5 times in order to reduce the effects of such factors as the processor, internet speed and so on. Examples of time required for data obtain are given in Table 1.

INSERT TABLE 1 HERE

After selection, a csv file containing the data is written in hard drive. 257 different user data were generated for the selected filters via Analytics. The created file is 0.0078 MB in size. The sample records of the created file are as shown in Table 2.

INSERT TABLE 2 HERE

3.2. Access Logs as A Data Source

This part of the work was designed to make a comparison with the proposed model. In this chapter, necessary and similar conversion are made to reach the data obtained by using the Analytics. The consequences are discussed in next chapters.

The log file that the web server for the relevant period has been downloaded from the server. This downloaded file contains 1,695,858 rows and file is 442.1 MB in size. On this file, the data cleaning process has been done with the methods suggested in the literature.

User Identification: For user identification, IP and browser agent name are used. However, the referrers' page control is not used because of the improved menu structure let every page access to another. In this merging process, the IP address and browser header names are combined with the md5 hash function to provide a quick search.

Session identification**:** The access log of every user is split into sessions. Time- oriented heuristic based session period algorithms are used for session identification. At this point, if a user has accessed a web page for more than 30 minutes, then this session will be separated into more than one session (Sukumar et al., 2016). For this filtering "30 min" value used for identification.

Data Cleaning: Web usage mining is used to mine the user access behavior and for this reason, all log records relevant to non-human access have to be removed from the weblog (Sukumar et al., 2016). The agent name values are used in the detection of these bots. Another filtering is done for media links and other sources. All media, style and script files have been removed from log data.

In this method, the 01-23.07.2018 range was used. After this pre-processing process, data for 927 users were created. The elapsed times are shown in Table 3 by running the joining operation 5 times. The resulting data file is 40,405 bytes in size. Data pre-processing timing examples are as follows:

INSERT TABLE 3 HERE

3.3. Comparison of Methods

There are different results when comparing data obtained with Analytics and web log. The number of different users for the same period in Yandex Metrica, another analysis tool, is 279. This suggests that the analysis from the server logs has produced incorrect results. In addition, the additional data that can be obtained from Analytics provides much better analysis.

Table 4 compares the efficiency of analysis studies between server logs and developed software with integrated Analytics.

INSERT TABLE 4 HERE

It can be seen that the integrated method seriously reduces data cleaning and preprocessing processes. These earned times can be used in the modeling and analysis process, which are create real benefits for. To demonstrate the completeness of the mining process, the data obtained with the application developed with the Analytics integration has been clustered with the K-means algorithm. example of this model is given in Figure 4 below.

INSERT FIGURE 4 HERE

4. Result and Discussion

Considering the developing technology of the web usage mining of the application developed by the study, it has been tried to show benefits to the new data sources and to the new integration. In this context, as a result of the work done with both methods, only about 9% of the server records give useful information and the remaining records have to be excluded. Even this situation cannot provide sufficient separation of users. The use of modern mobile internet, the use of multiple devices is a disadvantage only for analysis with server records.

With the proposed model and the developed application, the creation of a new data set, the compilation of the data and the analysis on this data have been abbreviated with the time that can be measured by the minute. Continuation of this study is supported by different data types and data mining methods for e-commerce sites, and Recommender System development and automation are planned.

References

Clifton, B. (2012). *Advanced web metrics with Google Analytics*: John Wiley & Sons.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems, 1*(1), 5-32.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media.

Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management, 32*(3), 477-481.

Reddy, K. S., Reddy, M. K., & Sitaramulu, V. (2013). *An effective data preprocessing method for Web Usage Mining.* Paper presented at the Information Communication and Embedded Systems (ICICES), 2013 International Conference on.

Seyrek, İ. H., & Ata, H. A. (2010). Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. *Journal of BRSA Banking & Financial Markets, 4*(2).

Sukumar, P., Robert, L., & Yuvaraj, S. (2016). *Review on modern Data Preprocessing techniques in Web usage mining (WUM).* Paper presented at the Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on.

Figures and Tables

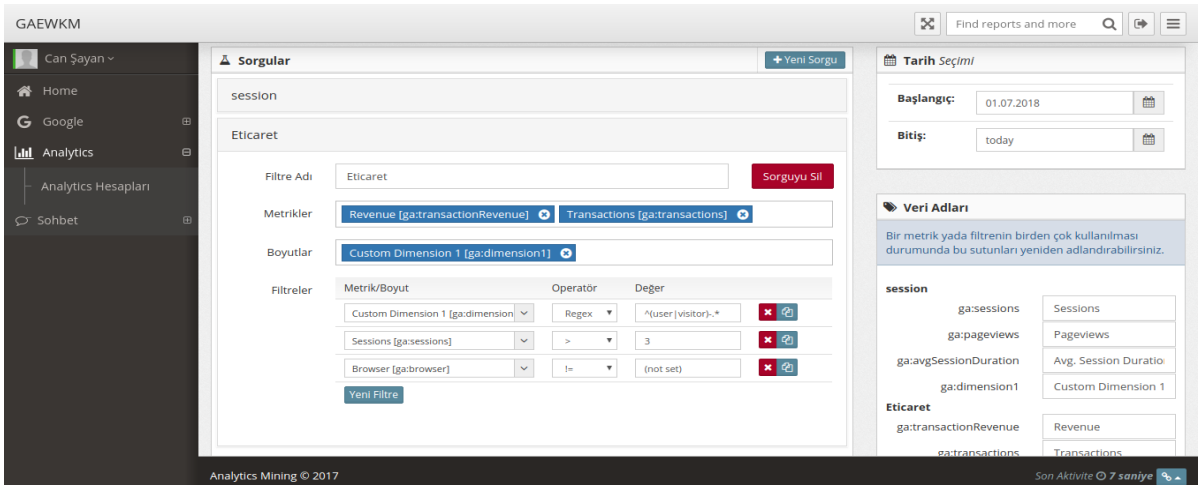

Figure 1. Data selection application interface

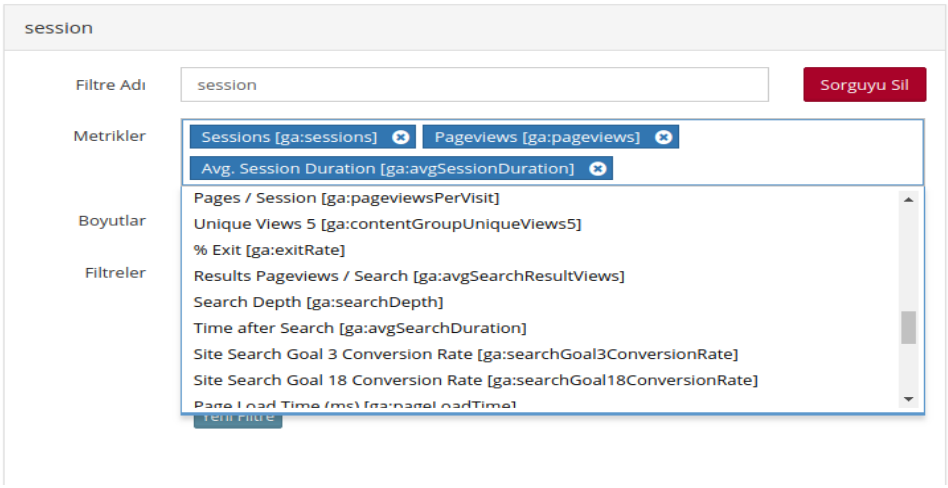

Figure 2. Property selection interface

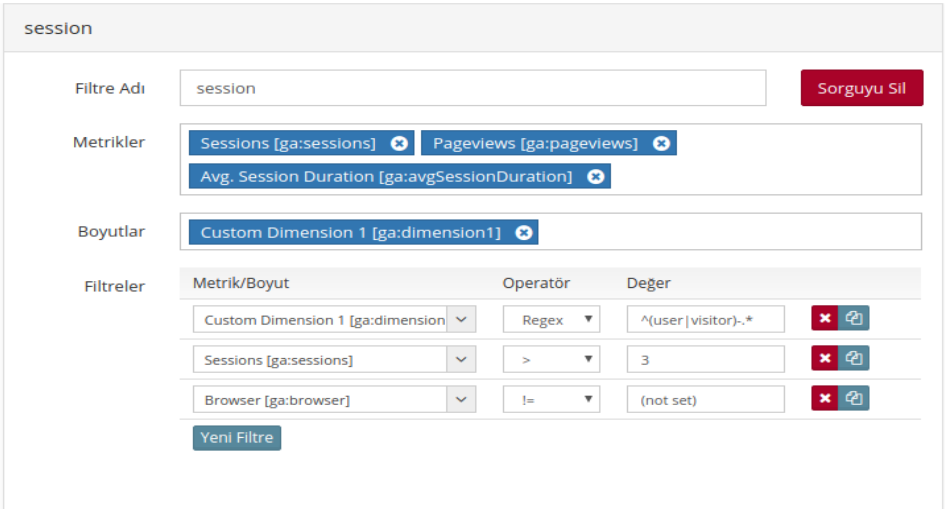

Figure 3. Criterion selection interface

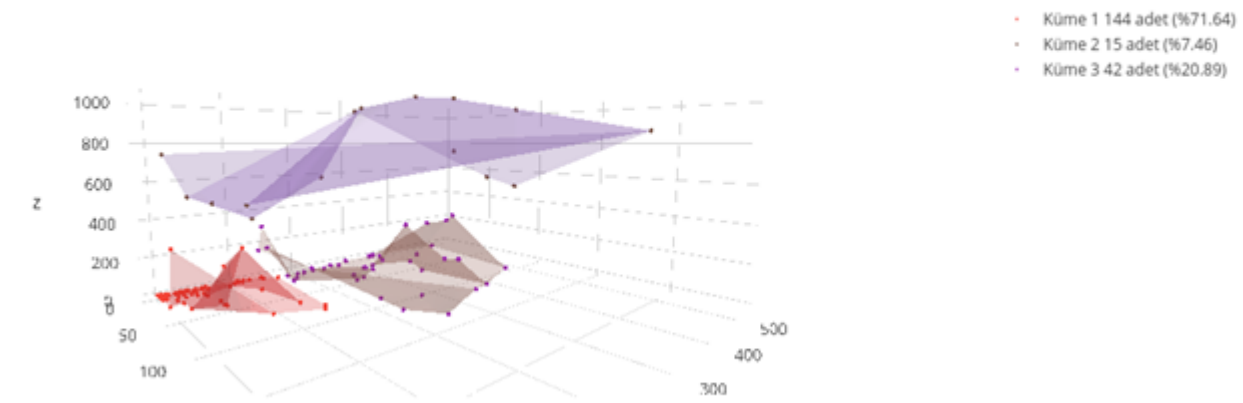

Figure 4. Pageview, average session duration and revenue comparison

Table 1.

Data preparation times for each run(Millisecond)-Proposed Model

Runs					Average
1.	2.	3.	4.	5.	
1282	598	581	583	583	725,4

Table 2.

Application record examples

dimension	transactionRevenue	transactions	sessions	pageviews	avgSession Duration
user-10220	0	0	77	1401	1959,76
user-10258	100	2	5	68	236,4
user-10326	0	0	5	17	600,4
user-10355	0	0	102	1206	1521.55

Table 3.

Data preparation times for each run(Millisecond)-Access Logs

Runs					Average
1.	2.	3.	4.	5.	
22929	22451	22261	23030	22469	22628

Table 4.

Comparison of Methods

	AccessLogs	Recommended Analytics Integrated Software
Records Count	1.695.858	257
Cleaned Records Count	1.529.094	0
Average Data Obtain Time	22628 millisecond	725 millisecond
Data Type	4	~699