Why Google Analytics Cannot Be Used For Educational Web Content

Sanda-Maria Dragoş
Chair of Computer Systems, Department of Computer Science
Faculty of Mathematics and Computer Science
"Babes-Bolyai" University 400084, Cluj-Napoca, Romania
Email: sanda@cs.ubbcluj.ro

***Abstract*—Current web analytics are focused on e-commerce sites, where the visits have to converge in a purchase. The behavior of e-learning environments users is driven by information acquiring. The learning process takes time, and therefore a visit on an educational site does not apply to the heuristics used by most analytics instruments (i.e. ending a visit after 30 minutes of inactivity). Moreover, an integrated analytics instrument may benefit from extended knowledge to better identify unique visitors. This paper proves that such a system is much more reliable than a system that bases its decisions on cookies (e.g., Google Analytics).**

***Index Terms*—web usage mining; web analytics; Google Analytics; e-learning;**

I. Introduction

Usage patterns are determined in order to better understand and provide for the needs of web users. Web usage mining [1] is a research area that uses data mining techniques to discover such patterns.

Web analytics is a part of web usage mining that has emerged in the corporate world, focusing on profitability in terms of how much money the web-site is making, macro and micro conversions [2]. Research has also been done [3] for using web usage mining to improve web-based learning instruments. Learners may benefit from new facilities such as automatically guidance, and recommendations of activities and resources that favors and improves the learning, while educators can be helped in assessing all activities performed by learners and evaluate the effectiveness of the structure of the course content on the learning process.

Although the e-commerce and e-learning have many similarities, their main objectives and techniques differ. While the objective of data mining in e-commerce is tangible and measurable (i.e., increasing profit) in e-learning it is more subjective and difficult to quantify (i.e., improve the learning).

II. Background

The most used data mining techniques are [4]: association rules, sequential patterns, clustering, classification and statistical analysis. Association rules imply discovery of relationships between the requested URLs. Sequential patterns are used to determine time ordered sequences of URLs that were accessed by users in order to predict future ones. Clustering and classification are both classification methods. Clustering is the process of unsupervised grouping of objects (i.e., users, events, sessions, pages), while classification is a supervised classification based on similar characteristics.

A. Web Analytics

According to the Web Analytics Association, Web Analytics (WA) is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing web usage [5]. The measurement of website traffic is done by using web analytics metrics.

The main bodies who have input in defining web analytics metrics are Jicwebs (Industry Committee for Web Standards)/ABCe (Auditing Bureau of Circulations electronic, UK and Europe), The WAA (Web Analytics Association, US) and to a lesser extent the IAB (Interactive Advertising Bureau). Both the ABCe and the WAA provide documents (i.e., a Report [6] and a Draft for Public Comment [5]) that contain lists of such definitions. The main metrics defined there are described below.

1) Page view / Page impression / Page request: - A request for a web page. However, a single page view may generate more file requests (images, *.js* and *.css* files) from the web server. Such files are irrelevant in the web analysis process.

2) Visit / Session: - In practice[1] both terms are used for visits, because it cannot be determined if a visitor viewed other pages from other domains. A *visit* is defined as a series of page requests from the same uniquely identified client with a time of no more than 30 minutes between each page request. It is also considered that a single page view does not constitute a visit or a session; it is a *"bounce"*.

3) (Unique) Visitor / Unique Browser / User: - The uniquely identified client generating requests on the web server (log analysis) or viewing pages (page tagging) within a defined time period (i.e., day, week, month). A *unique visitor* counts once within the timescale. A *visitor* can make multiple *visits*. Identification is made either to the visitor's computer (not the person) via cookie and/or IP+User Agent, or based on a registered account data if such information is available.

4) Bounce / Single page Visits: - A visit that consists of a single page view. *Bounces* usually indicate the website's failure to engage the visitor.

[1]In theory the session is more strict in the sense that it is limited to a specific time interval during which there may be no other requests for pages from other domains. Thus, a visit may contain more sessions.

5) Visit Duration: - Average amount of time that *visitors* spend on the site each time they visit. It is calculated as the difference between the time stamps of the fist page and the last pages accessed during that visit. The time spent by the user on the last page can not be determined. Thus, *bounces* are not considered for calculating this value.

6) Page Views per Visit / Visit Depth: - Average number of *page views* a *visitor* accesses during a *visit*.

7) Frequency / Visits per Unique Visitors: - Frequency measures how often *visitors visit* a website. It is calculated by dividing the total number of visits by the total number of unique visitors.

8) Recency: - Time since a *unique visitor* performed a specific action of interest (i.e., *visit* or any other event, e.g., download, use of a certain service, purchase, etc.).

9) Repeat Visitor / Repeat Unique Browser: - The number/percentage of *unique visitors* with more than one *visit* during the specified period of time.

These metrics can be grouped into three main classes [5]:

- **Building block terms** include the main metrics, *page views*, *visits* and *visitors* that make up the foundation for all web measurements. These metrics are also used as denominators formulas that determine all other metrics mentioned above.
- **Visit characterization terms** describe the behavior of a visitor during a website visit. Such metrics are *bounce*, *visit duration* and *page views per visit* and they are used to identify ways to improve a visitor's interaction with the web site.
- **Visitor characterization terms** contain metrics such as *frequency*, *recency* and *repeat visitors* that help distinguish website visitors. They enable segmentation of the visitor population to improve the accuracy and usefulness of analysis.

B. Google Analytics

Google Analytics (GA) [7] is a free web analytics instrument offered by Google, being the most widely used web analytics instrument [8].

Google Analytics is used by including a snippet of JavaScript code that the user adds into every page of his or her website. These *page tags* are used to collect the visitor data and send it back to Google data collection servers for processing. Also, GA sets first party cookies on each visitor's computer in order to be able to determine if the visitor has been to the site before.

The main GA limitations are characteristic to web analytics tools that collect on-site visitor data using page tagging:

- *Blocking JavaScript code*. This prevents traffic and users from being tracked, and leads to uncollected data.
- *Deleting or blocking cookies*. Such actions lead to inaccuracy as returning visitors cannot be tracked.

III. Main challenges

Data processing in web analytics starts with determining unique *visitors* and *visits*.

A. Unique user identification

One major challenge in web analytics is to identify unique visitors. One method is to identify them based on their IP addresses and the User Agent [6]. An alternative is to use cookies. Therefore, Google Analytics and other web analytics instruments use them to determine unique visitors. Cookies are used because IP addresses are not always unique to users and may be shared by large groups or proxies. However, there are other circumstances in which both of these (i.e., IP + User Agent and cookies) methods are inaccurate:

- **Multiple IP addresses - Single Visitor** - An individual that accesses the website from different locations/devices will have different IP addresses (respectively different cookie ID) from visit to visit and thus will be counted more than once. This makes tracking repeat visits from the same user difficult.
- **Multiple User Agents - Singe Visitor** - A user that uses more than one browser, even on the same machine, will appear as multiple users.

Moreover, cookies can be deleted or blocked.

The most accurate solution is to use registered user account information in order to identify individuals, especially for e-learning as most web-based educational systems use user authentication. This solution is the most realistic one and can be implemented only in an integrated system.

B. Visit/Session identification

Identifying accurate visits is not a trivial task. That is mainly because HTTP protocol is stateless and connectionless. Thus, it is virtually impossible to determine when a user is consulting the site or visiting other sites or if actually leaves the website. Moreover, some ISPs or privacy tools randomly assign each request from a user to one of several IP addresses. Although rare, in these cases, a single server session can have multiple IP addresses.

There are three main heuristics that are generally used to determine the visit termination:

1) temporal heuristics that restricts the duration of the entire visit to a predefined upper bound (usually accepted as 30 minutes) [9]
2) temporal heuristics that limits the time spent on any page to a threshold value accepted as 30 minutes according to [5], [6]
3) heuristic based on the navigational patterns where all pages within a session have to be linked directly or indirectly (i.e., the page has a referral that is a page accessed previously in the current visit).

The first two methods, although they might work on e-commerce sites, where the client makes the purchase and leaves the site, it is not applicable to e-learning as the learning process takes time.

IV. Integrated versus third party analysis

This paper presents the results obtained by interpreting the same web usage data by two different web analytics instruments. On one side there is an analytics instrument, called

WATEC (Web Analytics Tool for Educational Content) [10], that is integrated with an e-learning system called PULSE (Php Utility used in Laboratories for Student Evaluation) [11]. The benefit of this integration is the fact that WATEC can access student account information in order to better identify individuals. On the other side there is a Google Analytics-Like (GAL) instrument which uses cookies in order to identify unique visitors. GAL was built by the author specifically for the tests presented in this paper.

The aim of this study is to determine the differences between by the two approaches.

A. WATEC versus GAL

The time spent by a unique visitor on a webpage will be referred further on as the *time-on-page* and it is calculated through substraction of the access time of that page from the access time of next webpage within the same visit.

To compare the two instruments, they both analyze the web traffic on PULSE over a given period of time.

1) Data collection: A logging system records all PULSE accesses into a MySQL database. The collected information contains the following data fields:

- The time stamp of the request
- The IP address of the originating web page request
- Full request-URI, including the domain, the requested URL, and any applicable query parameters
- Full unmodified User-Agent string
- Referrer URL
- Student login ID (used only by WATEC)
- Cookie ID (used only by GAL)

This system records only webpages, and therefore no *data cleaning* is required.

2) Unique visitors: For visitor identification, WATEC uses the login ID of each student, while GAL uses cookie ID's.

3) WATEC visits: A WATEC visit consists of all web pages accessed consecutively from the same IP and User Agent (UA) by a WATEC *visitor* before login and until logout or closed browser. The pages accessed before login (usually only the login page) have the login field empty and they have to have the *time-on-page* less than 30 minutes. All other pages accessed while the *visitor* is authenticated have the login field containing the same ID and have no *time-on-page* restriction. The session concludes upon accessing the logout page, or if the IP address or UA changes, or the browser is closed. In normal circumstances it is impossible to determine when a user closes a browser only based on web access logs. However, upon closing the browser the PHP predefined variable SESSION resets, and thus the PULSE login session terminates and no login ID is recorded. Therefore, WATEC considers that any page access from the same IP+UA with an empty or different login ID field that follows another page access that has a login ID field is requested after a close browser event.

Formally a WATEC visit can be described as $W = [P_1, P_2, ..., P_n]$, where P_i is a page from the WATEC visit that has to satisfy the following conditions:

1) $\forall i$: $1 \leq i < n$, $T(P_i) \leq T(P_{i+1})$ (timestamp ordering)
2) $\forall i$: $1 \leq i < n$, $IP(P_i) = IP(P_{i+1})$ (same IP)
3) $\forall i$: $1 \leq i < n$, $UA(P_i) = UA(P_{i+1})$ (same UA)
4) $\exists b, 1 < b \leq n$ so that $\forall j$: $1 \leq j \leq b$, $EmptyLogin(P_j) = true$ and $\forall k$: $b < k \leq n$ $EmptyLogin(P_k) = false$ (before and after login)
5) $\forall j$: $1 \leq j < b$, $T(P_{j+1}) - T(P_j) < 1800$ (*time-on-page* for empty login)
6) $T(P_n) > T(P_{n+1})$ or $IP(P_n) \neq IP(P_{n+1})$ or $UA(P_n) \neq UA(P_{n+1})$ or $URL(P_n) \supset$ "$logout$" or ($EmptyLogin(P_n) = false$ and $Login(P_{n+1})! = Login(P_n)$) (terminate condition)

Here $T(P_i)$ is the timestamp of the page P_i representing the access time measured in the number of seconds since the Unix Epoch (i.e., January 1 1970 00:00:00 GMT).

After all visits were determined, WATEC eliminates the ones that do not have at least an entry with a login ID. Those visits do not represent valid visits, are not generated by an authenticated PULSE user.

4) GAL visits: A GAL visit contains all pages accessed consecutively with the same cookie ID that have the *time-on-page* less than 30 minutes.

The formally description of a GAL *visit* is $A = [P_1, P_2, ..., P_m]$, where all pages P_i have to satisfy the following conditions:

1) $\forall i$: $1 \leq i < m$, $T(P_i) \leq T(P_{i+1})$ (timestamp ordering)
2) $\forall i$: $1 \leq i < m$, $cookieID(P_i) = cookieID(P_{i+1})$ (same cookie ID)
3) $\forall i$: $1 \leq i < m$, $T(P_{i+1}) - T(P_i) < 1800$ (*time-on-page*)
4) $T(P_m) > T(P_{m+1})$ or $cookieID(P_m) \neq cookieID(P_{m+1})$ or $T(P_{m+1}) - T(P_m) >= 1800$ (terminate condition)

In order to better delimitate sessions the log data was clustered/grouped by cookie ID's and timestamp. Then, WATEC and GAL sessions were determined as described above.

B. Test Results

The listing presented in Fig. 1 depicts a sequence from the log data that shows differences between WATEC and GAL visits. Each line from these listings has the following fields:

- id from the IP addresses table
- id from the User Agents table (in light gray color)
- the date of the access (format: year-month-day)
- the time of the access (format: hour:minutes:seconds)
- *time-on-page* (computed using the time stamp field)
- Referral page URL (truncated for better visualization)
- access page URL (truncated for better visualization)
- login
- cookie ID

The thick line marks the end of the GAL visit. The GAL visit duration is specified above this line, as well as the terminate condition. WATEC visit termination is marked by a thin line, above which are specified the WATEC visit duration as well as the terminate conditions.

Most web analytics instruments consider that the visit ends after 30 minutes of inactivity. The listing from Fig. 1, however,

7108 1385 2011-04-14 19:21:01	**2"**	PULSE.php?2010-2011/SO1&cursuri&SO_curs4.other	PULSE.php?2010-2011/SO1&cursuri	bpri0034	**1951**
7108 1385 2011-04-14 19:21:03	**1' 47"**	PULSE.php?2010-2011/SO1&cursuri	PULSE.php?2010-2011/SO1&cursuri&SO_curs3.other	bpri0034	**1951**
7108 1385 2011-04-14 19:22:50	**36' 23"**	PULSE.php?2010-2011/SO1&cursuri&SO_curs4.other	PULSE.php?2010-2011/SO1&cursuri	bpri0034	**1951**
	3' 5"	**- GAL Time**			
7108 1385 2011-04-14 19:59:13	**4h 15' 25"**	PULSE.php?2010-2011/SO1&cursuri	PULSE.php?2010-2011/SO1&cursuri&SO_curs3.other	bpri0034	**1951**
	0"	**- GAL Time** - *BOUNCE* -			
	39' 28"	**- WATEC Close**			
7108 1385 2011-04-15 00:14:38	**35"**	PULSE.php?2010-2011/SO1&cursuri&SO_curs4.other	PULSE.php?2010-2011/SO1&cursuri		**2007**
7108 1385 2011-04-15 00:15:13	**6"**	PULSE.php?2010-2011/SO1&cursuri	PULSE.php	bpri0034	**2007**

Fig. 1. Sequence from the Log Data that Exemplifies the Different Interpretation between the WATEC Visit versus the GAL Visits.

proves that a HTTP session does not end after 30 minutes of inactivity. Thus, the student is still authenticated after a *time-on-page* of more that 36 minutes. In such circumstances, GAL considers the visit terminated, while WATEC which due to its integration with PULSE can access more detailed information for each page access (i.e., login name) will be able to determine the correct length of a visit, and end it (in this very case) only when the student closes the browser and thus looses his/her PULSE authentication as depicted on the penultimate line of the listing in Fig. 1.

Another difference in visit interpretation between the two analytics instruments is when the same individual uses more computers (i.e., different laboratories, at work, at home). In such circumstances GAL considers distinct visitors for each computer and all their visits as performed by different visitors. As an integrated instrument, WATEC is always able to determine the identity of an user based on his/her login name, and therefore such misleading interpretations are eliminated.

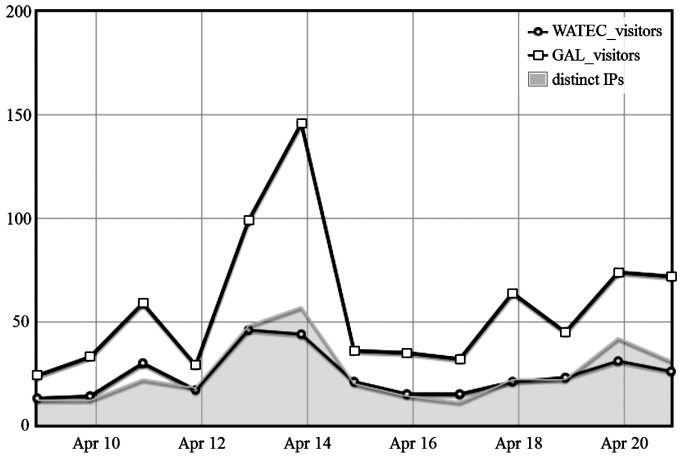
The graphics in Fig. 2 depict WATEC's and GAL's comparative evolution in terms of the number of visits and visitors. The time interval is between the 9th and the 21st of April 2011. The filled area marks the number of distinct IP's recorded over that period.

As shown in Fig. 2(a), GAL identified more *visitors* that WATEC. The discrepancies are due to GAL's difficulty to accurately identify visitors. The challenges in identifying unique individuals arise mainly because a single student may access the website from multiple locations (e.g., school, home) and/or by using different browsers. In such circumstances GAL cannot distinguish between different individuals, unlike WATEC which uses PULSE's authentication information. Moreover, cookies can be deleted or denied and therefore even if a user revisited (see last two columns in Fig. 1), GAL considers him/her to be a new visitor.

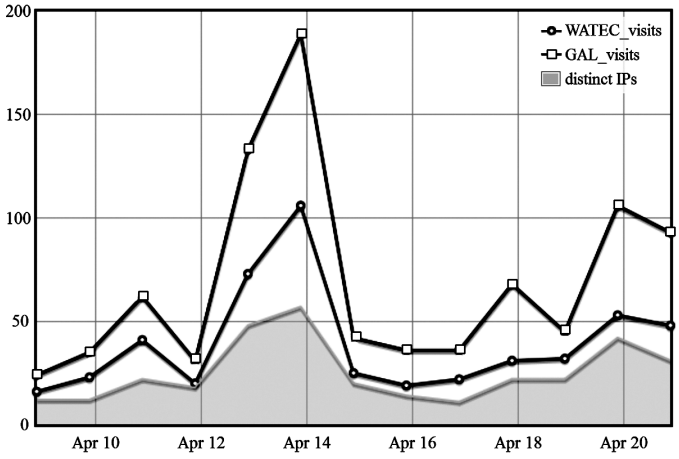
Fig. 2(b) shows that GAL determines more *visits* than WATEC although they both process the same web usage log information. The main challenge in identifying a visits consists in determining when it ends. Because HTTP is a stateless and connectionless protocol, it is virtually impossible to determine when a user actually leaves a website. Thus, GAL applies the heuristics used by most web analytics instruments by considering the visit to end after a 30 minutes period of inactivity. However, the learning process takes time and these tests have proven that a HTTP session does not end after 30 minutes of inactivity because a PULSE user will still be authenticated even after more 30 minutes of inactivity.

The cases where the number of WATEC visits is less than the number of IP addresses, in Fig. 2(b), are due to the fact that WATEC does not consider non authenticated visits.

The values obtained for all metrics described in Section II-A for the time period between the 9th and the 21st of April 2011 are presented in Table I.



(a) Number of Visitors



(b) Number of Visits

Fig. 2. WATEC's and GAL's Comparative Results in Terms of *Visitors* and *Visits* for the Time Period between the 9th and the 21st of April 2011

The metrics that make up the foundation of all web measurements (i.e., the *building block terms*) are:

- *page views:* Having the same log data the number of *page views* is the same in both interpretations and it is basically the number of entries in the log database.
- *visitors:* Distinct interpretations between WATEC and GAL in terms of *visitors* prove that there are more cookie IDs than actual PULSE users. The numbers presented in Table I show that for every PULSE user there are on average 11 cookies. This may be either because each individual uses different computer/devices and operating systems/browsers or because cookies are denied or deleted.
- *visits:* GAL counts almost 1.8 more *visits* than WATEC mostly because on the *time-on-page* restriction. The fact that WATEC does not take into account the unauthenticated visits is the other reason for the difference between the number of GAL's and WATEC's visits.

The *visit characterization terms* describe the behavior of visitors during their visits, so that:

- *bounce:* A high *bounce* rate can mean either that wrong people visit the site or that the site is poorly constructed. In the case presented in Table I, the explanation for the *bounce* visits is twofold. On one hand there are the uncounted WATEC's unauthenticated visits (i.e., individuals that landed on site by accident and not having a PULSE account, and therefore no reason to stay or unhuman traffic). Many of such visits can be bounces. On the other hand there are GAL's fragmented visits due to its *time-on-page* constraint. Some of such fragments are bounces as exemplified in Fig. 1. However, GAL's bounce value is significant as it says that a quoter of its visits are bounces.
- *visit depth* and *visit duration:* Differences between GAL's and WATEC's *visit depth* and *visit duration* are also due to the *time-on-page* restriction which generates visit fragmentation for GAL.

All metrics that help distinguishing website visitors (i.e., *visitor characterization terms*) have higher WATEC values as it can determine more reliably if the same users return (even from different computers and/or browsers).

The most important metrics that can help determine the effectiveness of PULSE are *visit depth*, *visit duration*, *frequency* and *recency*. One single value (as provided in Table I) does not offer enough insight. The distribution of these metrics may help to understand what the median is and where are PULSE's outliers.

The distributions of the metrics that provide a sense of the way the content is consumed are presented in Fig. 3. In Fig. 3(a) it is presented the distribution of number of pages in each visit to PULSE (i.e., the *visit depth*) during the given time period. GAL determines that more than 33% of its visits are a single page views (e.g., *bounces*), while 30% of WATEC's visits contain 12 or more page views. Also, 58% of WATEC's visits contain 7 or more page views compared with 62% GAL's visits that contain 6 or less page views.

TABLE I
WEB ANALYTICS RESULTS FOR WATEC AND GAL FOR THE TIME PERIOD BETWEEN THE 9TH AND THE 21ST OF APRIL 2011

		WATEC	GAL
Building Block Terms	**Page views**	5663	5663
	Visitors	67	740
	Visits	509	900
Visit Characterization Terms	**Bounce**	0 %	25.25 %
	Visit depth (Page views per Visits)	11.13	6.29
	Visit duration	2h 8' 35"	6' 54"
Visitor Characterization Terms	**Frequency** (Visits per Visitor)	7.6	1.22
	Recency	1.71 days	0.01 days
	Repeat visitors	86.84 %	17.78 %

Pageviews in the visit	Visits with this many pageviews		Percentage of all visits	
	WATEC	GAL	WATEC	GAL
1 pageviews		304	0	33.78
2 pageviews	4	40	0.79	4.44
3 pageviews	37	64	7.27	7.11
4 pageviews	42	56	8.25	6.22
5 pageviews	73	94	14.34	10.44
6 pageviews	59	64	11.59	7.11
7 pageviews	40	37	7.86	4.11
8 pageviews	28	35	5.5	3.89
9 pageviews	29	26	5.7	2.89
10 pageviews	26	25	5.11	2.78
11 pageviews	18	13	3.54	1.44
12 + pageviews	153	142	30.06	15.78

(a) Visit depth

Duration of visit	Visits with this duration		Percentage of all visits	
	WATEC	GAL	WATEC	GAL
0-10 seconds	15	344	2.95	38.22
11-30 seconds	72	120	14.15	13.33
31-60 seconds	48	59	9.43	6.56
61-180 seconds	89	112	17.49	12.44
181-600 seconds	69	102	13.56	11.33
601-1800 seconds	85	101	16.7	11.22
1801 + seconds	131	62	25.74	6.89

(b) Visit duration

Fig. 3. Distributions on *Visit* Characterization Terms on the Time Period Between the 9th and the 21st of April 2011

Fig. 3(b) presents the *visit length* (as temporal duration) distribution over the same period of time. This metric denotes the quality of the visit. Due to GAL's visit fragmentation, there are more that 38% of visits that lasted less that 10 seconds. More that 25% of WATEC's visits lasted more that 30 minutes compared with 7% of GAL's visits that lasted more that 30 minutes.

It is crucial for any type of content site to get a sense for how strongly attached are the *visitors* to the site.

Fig. 4(a) shows the *frequency* with which a visitor returned to PULSE. This metric represents the loyalty of PULSE users. Ideally, all PULSE users have to revisit the site at least once a

week as the content on PULSE changes weekly. The recorded period is of 13 days, so most students are expected to visit the site twice. Most GAL visitors (82%) visited only once. However, most of WATEC visitors (75%) visited at least 3 times, while 87% of visitors visited at least 2 times.

Count of visits from this visitor	Visits that were the visitor's nth visit		Percentage of all visits	
	WATEC	GAL	WATEC	GAL
1 times	67	740	13.16	82.22
2 times	61	127	11.98	14.11
3 times	56	25	11	2.78
4 times	50	4	9.82	0.44
5 times	44	1	8.64	0.11
6 times	41	1	8.06	0.11
7 times	38	1	7.47	0.11
8 times	28	1	5.5	0.11
9 times	23		4.52	0
10 + times	101		19.84	0

(a) Frequency

Previous visit was tracked	Total visits by period		Percentage of all visits	
	WATEC	GAL	WATEC	GAL
First visit	67	740	13.16	82.22
Same day	205	157	40.28	17.44
1 days ago	71	1	13.95	0.11
2 days ago	30	1	5.89	0.11
3 days ago	29		5.7	0
4 days ago	15		2.95	0
5 days ago	21	1	4.13	0.11
6 days ago	43		8.45	0
7 days ago	16		3.14	0
8-14 days ago	12		2.36	0

(b) Recency

Fig. 4. Distributions on *Visitor* Characterization Terms on the Time Period Between the 9th and the 21st of April 2011

The interval of time since visitors last visit is called *recency* and its distribution over the recorded period of time is presented in Fig. 4(b). It is expected that every visitor has to revisit the site weekly in order to consult the changes on PULSE. The result in Fig. 4(b) show that within 6 days recency 100% of GAL's visitors revisited PULSE compared with 95% of WATEC's visitors. Although satisfactory, WATEC's performance is worse in this case than GAL's. The explanation lays in the fact that GAL's cookies can expire, can be denied or deleted. Thus, GAL has a difficulty in keeping track of *repeat visitors* as it can also be seen in the last row of Table I.

V. Conclusion and Future Work

This paper shows that integrated analytics are more accurate and therefore more reliable than legacy analytics which in this case is represented by a Google Analytics-like instrument. Even when using the same web usage log data, the differences between the two techniques are significant. This discrepancies are even more acute as the analysis in done on a e-learning system which is a content-based website, where the rules of e-commerce, for which Google Analytics is mainly used, do not apply entirely.

The tests presented in this paper showed that GAL's interpretation lead to determining more visits, but 25% of them are *bounces* and the rest of them are much shorter (as number of pages viewed and duration) as they really are. Also, GAL determined that even if are plenty of visitors, they do not visit very often and almost all of their visits are in the same day.

WATEC on the other hand offers a more realistic and therefore reliable interpretation by using the login information.

As a parallel research, the author studied in [12] the differences between WATEC and a web analytics instrument that identifies visitors based on the IP and user agent (UA). The results obtained by this instrument, although worst than WATEC (as it was expected) were better than GAL's results.

As a future work, the author wants to focus on the content most people are consuming on PULSE by segmenting the data and thus measure the effectiveness of individual pages. Then, the natural step is to move from getting the insights gained here and those which will be obtained in the future to make improvements on PULSE suggested by this knowledge. Also, critical information is gained when web analytics is combined with other related instruments such as semantic web. Therefore, the author is already studying this new approach.

Acknowledgment

This material is partially supported by the Romanian National University Research Council under award PN-II IDEI 2412/2009.

References

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, January 2000.

[2] D. Waisberg and A. Kaushik, "Web analytics 2.0: Empowering customer centricity," *SEMJ.org*, vol. 2, no. 1, 2009.

[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, p. 135146, July 2007.

[4] N. K. Tyagi, A. K. Solanki, and S. Tyagi, "An algorithmic approach to data preprocessing in web usage mining," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 279–283, July-December 2010.

[5] Web Analytics Association, "Web analytics definitions," Web Analytics Association, Draft for Public Comment, 2008.

[6] The Joint Industry Committee for Web Standards (JICWEBS), "Reporting standards. website traffic." Auditing Bureau of Circulations electronic (ABCe), Report 1, 2011.

[7] Google, "Google analytics," http://www.google.com/analytics/, viewed on May 2011.

[8] W3Techs, "Usage statistics and market share of traffic analysis tools for websites," http://w3techs.com/technologies/overview/traffic_analysis/all, viewed on May 2011.

[9] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in web-usage analysis," *INFORMS Journal on Computing*, vol. 15.

[10] S. Dragos and R. Dragos, "WATEC: a Web Analytics Tool for Educational Content," *KEPT 2009*, vol. Selected Papers, pp. 320–327, 2009.

[11] S. Dragos, "PULSE - a PHP Utility used in Laboratories for Student Evaluation," in *International Conference on Informatics Education Europe II (IEEII)*, Thessaloniki, Greece, November 2007, pp. 306–314.

[12] ——, "Why integrated e-learning analytics are the best solution?" to be presented in the International Conference on Intelligent Computer Communication and Processing (ICCP), to be held August 25–27, 2011 in Cluj-Napoca, Romania.