

# Why integrated E-learning Analytics are the best solution?

Sanda-Maria Dragos

Faculty of Mathematics and Computer Science,  
"Babes-Bolyai" University, Cluj-Napoca, Romania  
Email: sanda@cs.ubbcluj.ro

**Abstract**—Visitors and visits are building block concepts for most web analytics instruments. The way a visits or a unique visitor are determined is crucial for interpreting web log data. Tests show that the most accurate analytics method (i.e., using authentication data) have to be used on educational content to obtain relevant insights. Most current web analytics are focused on e-commerce sites where the objective is quantifiable (i.e., purchases) as opposed to e-learning where the objective is teaching.

## I. INTRODUCTION

With the emergence of web-based applications (e.g., e-commerce, e-learning, on-line collaboration), analyzing and improving web communication has become essential in order to fulfill user expectations. Web usage mining is a research area that uses data mining techniques to discover usage patterns from web data in order to understand and provide for the needs of web users by creating and modifying website content [1]. 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], [4] 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

Data mining techniques may be applied only on a suitable shaped data that is obtained through the process of *preprocessing* [5]. This action implies a series of tasks [5], [4]:

- *Data cleaning* is the process of removing all data recorded in web logs that are useless for the mining process (e.g., graphics, scripts, css).
- *User identification* is the process of associating log entries to a connected user.

- *Session/visit identification* consisting of all logged contiguous actions performed by a user during a visit.
- *Data transformation and enrichment* consists in calculating new attributes from existing ones, respectively, considering other sources of data such as structure and/or content information in order to enrich the web log data. The latter can be done by using techniques as web structure mining, web content mining or semantic web [5].
- *Data formatting* which is used for a better access during the mining process. Solutions presented in [5] use relational databases or tree representations (e.g., WAP-tree, signature tree) for that purpose.

After the preprocessing phase, which can be very complex and time-consuming, pattern discovery techniques can be applied to data. There are multiple methods, from diverse fields such as statistics, data mining, pattern recognition, that can be applied on web usage data.

### A. Data Mining Techniques

The most used data mining techniques are [5], [3], [6]: association rules, sequential patterns, clustering and classification. 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.

### B. Web Analytics

Web analytics is based<sup>1</sup> on statistical interpretation of some metrics meant to describe the performance of a site.

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 [8] and a Draft for Public Comment [7]) that contain lists of such definitions. The main metrics defined there are:

<sup>1</sup>According to the Web Analytics Association [7], Web Analytics (WA) is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing web usage.

- **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 process of web usage mining (thus, are excluded from the logged data during the *data cleaning* process as mentioned Section II) or during web analysis (see “*Data filtering and exclusions*” in [8]).
- **Visit / Session** In practice<sup>2</sup> both terms are used for visits. A *visits* 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*”).
- **(Unique) Visitor / Unique Browser / Unique 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.
- **Bounce / Single page Visits** A visit that consists of a single page view. *Bounces* usually indicate the website’s failure to engage the visitor.
- **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 first 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.
- **Page Views per Visit / Visit Depth** Average number of *page views* a *visitor* accesses during a *visit*.
- **Frequency / Visits per Unique** 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.
- **Recency** Time since a *unique visitor* performed a specific action of interest (i.e., *visit* or any other event, e.g., purchase, download, use of a certain service, etc.).
- **Repeat Visitor / Repeat Unique Browser** The number (or percentage) of *unique visitors* with more than one *visit* during the specified period of time.

### III. MAIN CHALLENGES

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

#### A. Unique user identification

One major difficulty in web usage mining as well as in web analytics is to identify unique visitors. One method is to identify them based on their IP addresses and the User

Agent [8]. However, there are many circumstances in which such assumption is inaccurate:

- **Single IP address - Multiple Visitors** All students use the same set of computers provided in laboratories. Thus, more students use the same computer and the same browser. Also, all users behind a proxy server will have the same IP address and therefore they may be seen as a single visitor.
- **Multiple IP addresses - Single Visitor** An individual that accessed the website from different locations/devices will have different IP addresses 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 - Single Visitor** A user that uses more than one browser, even on the same machine, will appear as multiple users.

The alternative is to use cookies. However, cookies can be deleted or denied.

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 [7], [8]
- 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 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], [11], that is

<sup>2</sup>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.

integrated with an e-learning system called PULSE (Php Utility used in Laboratories for Student Evaluation) [12], [13], [14]. 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 web analytics instrument named AIU (Analytics by IP address and User Agent) that uses the IP address and the User Agent in order to identify unique visitors. The aim of this study is to determine the differences between the two approaches.

#### A. WATEC versus AIU

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 subtraction of the access time of that page from the access time of previous webpage within the same visit.

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

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

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 AIU considers each pair of unique (IP, User-Agent) as being a unique visitor.

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  $W = [P_1, P_2, \dots, P_n]$ , 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 User Agent)

- 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  $(EmptyLogin(P_n) = false \text{ and } Login(P_{n+1})! = Login(P_n))$  (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) *AIU visits*: An AIU visit contains all pages accessed consecutively from the same IP, using the same User Agent (UA) by a AIU *visitor* that have the *time-on-page* less than 30 minutes.

The formally description of an AIU *visit* is  $A = [P_1, P_2, \dots, P_m]$ , 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)  $\forall i: 1 \leq i < m, IP(P_i) = IP(P_{i+1})$  (same IP)
- 3)  $\forall i: 1 \leq i < m, UA(P_i) = UA(P_{i+1})$  (same User Agent)
- 4)  $\forall i: 1 \leq i < m, T(P_{i+1}) - T(P_i) < 1800$  (*time-on-page*)
- 5)  $T(P_m) > T(P_{m+1})$  or  $IP(P_m) \neq IP(P_{m+1})$  or  $UA(P_m) \neq UA(P_{m+1})$  or  $T(P_{m+1}) - T(P_m) >= 1800$  (terminate condition)

In order to better delimitate sessions the log data was clustered/grouped by IP addresses, User Agents and timestamp. Then, WATEC and AIU sessions were determined as described above.

#### B. Test results

Listings presented in Figures 1 and 2 depict sequences from log data that show differences between WATEC and AIU 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* which is computed from the time of access
- Referral page URL (truncated for better visualization)
- access page URL (truncated for better visualization)
- login

The thick line marks the end of the AIU visit. The AIU visit duration is specified above this line, as well as the terminate condition. WATEC visit terminations are 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 a visit ends after 30 minutes of inactivity. The listing from Figure 1 proves

|                          |     |                     |         |                                                                     |          |
|--------------------------|-----|---------------------|---------|---------------------------------------------------------------------|----------|
| 3156                     | 551 | 2010-10-20 21:19:47 | 31"     | /~sanda PULSE.php                                                   |          |
| 3156                     | 551 | 2010-10-20 21:20:18 | 5"      | PULSE.php PULSE.php                                                 | rdir0932 |
| 3156                     | 551 | 2010-10-20 21:20:23 | 58"     | PULSE.php PULSE.php?2010-2011/AC                                    | rdir0932 |
| 3156                     | 551 | 2010-10-20 21:21:21 | 1' 53"  | PULSE.php?2010-2011/AC PULSE.php?2010-2011/AC&lab1                  | rdir0932 |
| 3156                     | 551 | 2010-10-20 21:23:14 | 6"      | PULSE.php PULSE.php?2010-2011/AC                                    | rdir0932 |
| 3156                     | 551 | 2010-10-20 21:23:20 | 49' 46" | PULSE.php?2010-2011/AC PULSE.php?2010-2011/AC&assignment&1&rdir0932 | rdir0932 |
| 3' 53" - AIU Time        |     |                     |         |                                                                     |          |
| 3156                     | 551 | 2010-10-20 22:13:06 | 3"      | PULSE.php PULSE.php?2010-2011/AC                                    | rdir0932 |
| 3156                     | 551 | 2010-10-20 22:13:09 | 20' 19" | PULSE.php?2010-2011/AC PULSE.php?2010-2011/AC&lab1                  | rdir0932 |
| 3156                     | 551 | 2010-10-20 22:33:28 | 3"      | PULSE.php?2010-2011/AC&lab1 PULSE.php?2010-2011/AC&logout           |          |
| 1h 14' 1" - WATEC Logout |     |                     |         |                                                                     |          |

Fig. 1. WATEC visit versus AIU visit. AIU's *time-on-page* condition

however 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 49 minutes. In such circumstances, AIU 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 logged out as depicted on the last line of the listing in Figure 1.

Another difference in visit interpretation between the two analytics instruments occurs when more PULSE users use the same computer. In this case AIU, which uses IP+UA to determine new visitors counts those visits as coming from the same individual. The results shown in Figure 2 prove that very assumption. The first visit ends with logout which was determined by WATEC. The second visit ends with closing the browser, which results in loosing the login name. Thus, the first page access from the very next session from the same IP and User Agent (all records were sorted by IP, User agent and access timestamp) has no login name as the next user did not logged in yet. AIU, however, cannot sense these changes and therefore considers all three visits as a single one, although

they were visits of distinct PULSE users as it can be observed in the last column of the listing from Figure 2.

There are also the circumstances when the same individual uses more computers (i.e., different laboratories, at work, at home) in which AIU 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 Figure 3 depict WATEC's and AIU's comparative evolution in terms of the number of visits and visitors. The time interval is between 1 and 30 March 2011. The yellow filled area marks the number of distinct IP's recorded over that period.

As shown in Figure 3(a), AIU determined more visits that WATEC. There is one exception, in 20 March 2011. Such cases occur in the circumstances described by Figure 2, when more users use the same computer and browser. These results show however that there are more cases (like those depicted in Figure 1) when AIU considers a visit ended after each *time-on-page* > 30 minutes. Figure 3(b) shows that students use multiple computers, and therefore WATEC has less unique

|                         |     |                     |               |                                                                            |          |
|-------------------------|-----|---------------------|---------------|----------------------------------------------------------------------------|----------|
| 3101                    | 340 | 2010-10-20 18:40:17 | 17' 15"       | PULSE.php?2010-2011/PW&logout PULSE.php?2010-2011/PW                       |          |
| 3101                    | 340 | 2010-10-20 18:57:32 | 4"            | PULSE.php?2010-2011/PW PULSE.php                                           | tlir0974 |
| 3101                    | 340 | 2010-10-20 18:57:36 | 8' 44"        | PULSE.php PULSE.php?2010-2011/PW&assignment&3&tlir0974                     | tlir0974 |
| 3101                    | 340 | 2010-10-20 19:06:20 | 2"            | PULSE.php?2010-2011/PW&assignment&3&tlir0974 PULSE.php?2010-2011/PW&logout |          |
| 26' 3" - WATEC Logout   |     |                     |               |                                                                            |          |
| 3101                    | 340 | 2010-10-20 19:06:22 | 2' 9"         | PULSE.php?2010-2011/PW&logout PULSE.php?2010-2011/PW                       |          |
| 3101                    | 340 | 2010-10-20 19:08:31 | 6"            | PULSE.php?2010-2011/PW PULSE.php                                           | saie0899 |
| 3101                    | 340 | 2010-10-20 ...      | ...           | ...                                                                        | saie0899 |
| 3101                    | 340 | 2010-10-20 19:09:09 | 21"           | PULSE.php?2010-2011/WP PULSE.php?2010-2011/WP&assignment&4&saie0899        | saie0899 |
| 3' 8" - WATEC Close     |     |                     |               |                                                                            |          |
| 3101                    | 340 | 2010-10-20 19:09:31 | 12"           | PULSE.php?2010-2011/WP&logout PULSE.php?2010-2011/WP                       |          |
| 3101                    | 340 | 2010-10-20 19:09:43 | 7"            | PULSE.php?2010-2011/WP PULSE.php                                           | pmie0872 |
| 3101                    | 340 | 2010-10-20 19:09:50 | 128' 59" 090" | PULSE.php PULSE.php?2010-2011/WP&assignment&4&pmie0872                     | pmie0872 |
| 29' 33" - AIU IP+UA     |     |                     |               |                                                                            |          |
| 19" - WATEC IP UA Close |     |                     |               |                                                                            |          |

Fig. 2. WATEC visit versus AIU visit. WATEC's *visitor logout* condition and *browser close* condition

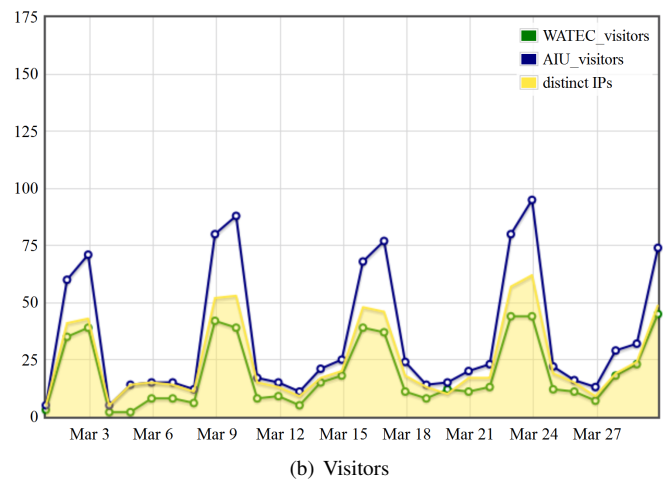
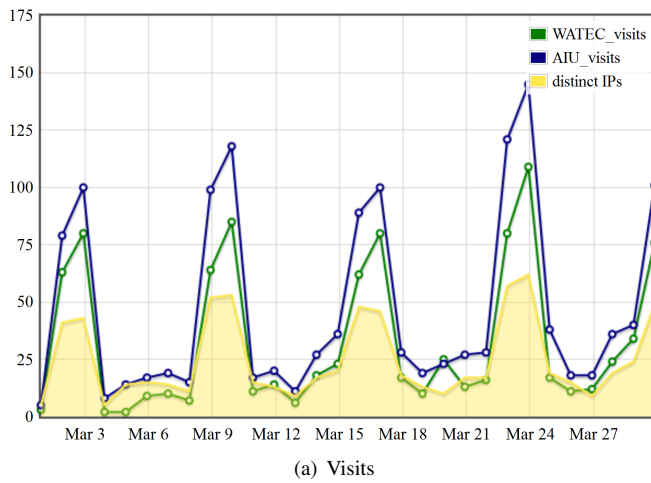


Fig. 3. WATEC's and AIU's comparative results for the time period between 1 March 2011 and 30 March 2011

visitors (fewer even than distinct IP addresses). The cases (Fig 3(a)) in which the number of WATEC visits is less than IP addresses are due to the fact that WATEC does not consider non authenticated visits.

*Page views*, *visits* and *visitors* are building block metrics as they make up the foundation for all web measurements be it web usage mining or web analytics. They are also used to determine other metrics like those that characterize visits (i.e., *bounce*, *visit depth*, *visit duration*) or metrics that characterize visitors behavior (i.e., *frequency*, *recency*, *repeat visitors*). The values obtained for the period 1-30 March 2011 are presented in Table I.

|                                          | WATEC     | AIU      |
|------------------------------------------|-----------|----------|
| <b>Page views</b>                        | 11675     | 11675    |
| <b>Visitors</b>                          | 92        | 584      |
| <b>Visits</b>                            | 978       | 1410     |
| <b>Bounce</b>                            | 0.51 %    | 22.99 %  |
| <b>Visit depth=Page views per Visits</b> | 11.94     | 8.28     |
| <b>Visit duration</b>                    | 24' 57"   | 8' 56"   |
| <b>Frequency = Visits per Visitor</b>    | 10.63     | 2.41     |
| <b>Recency</b>                           | 5.01 days | 2.7 days |
| <b>Repeat visitors</b>                   | 90.59     | 58.58    |

TABLE I  
WEB ANALYTICS RESULTS FOR WATEC AND AIU FOR THE TIME PERIOD  
BETWEEN 1 MARCH 2011 AND 30 MARCH 2011

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. Distinct interpretations between AIU and WATEC in terms of *visitors* prove that there are more (IP+UA)s used than actual PULSE users. The numbers presented above prove that every PULSE user uses an average of 6 different computer/devices and/or operating systems/browsers. AIU counts almost 1.5 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 another reason for the difference between the number of AIU's and WATEC's visits. The explanation for the *bounce*

visits is twofold. On one hand there is 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). Many of such visits can be bounces. On the other hand there are AIU fragmented visits due to the *time-on-page* constraint. Some of such fragments are bounces. Differences between AIU's and WATEC's *visit depth* and *visit duration* are also due to the *time-on-page* restriction which generates visit fragmentation. All metrics characterizing visitor behavior have higher WATEC values as it can determine more reliably if the same users return (even from different computers and/or browsers).

Figure 4 presents distributions for four of the metrics presented above (i.e., *visit depth*, *visit duration*, *frequency* and *recency*) in order to have a more insightful perspective on them. Figure 4(a) shows the distribution of number of pages in each visit over the time period 1-30 March 2011. AIU determines that almost 30% of its visits are a single page views, which is due the AIU's visit fragmentation. However, there are cases in which AIU considers within the same visit the activity of distinct PULSE users (see Figure 2). Most of such cases are contained in the last row of Figure4(a). 30% of WATEC's visits contain 12 or more page views, while 60% of WATEC's visits contain 7 or more page views compared with 62% AIU's visits that contain 6 or less page views.

Figure 4(b) depicts the *visit length/duration* distribution over the same period of time. This metric denotes the quality of the visit. Also due to AIU's visit fragmentation, there are more than 30% of visits that lasted less than 10 seconds. More than 20% of WATEC's visits lasted more than 30 minutes compared with 9% of AIU's visits that lasted more than 30 minutes. Among those log visits of AIU are also those that contain the activity of distinct PULSE users (see Figure 2).

Figure 4(c) shows the *frequency* with which a visitor returned to PULSE. This metric represents the loyalty of PULSE users and within the recorded period of time that have to return more than 4 times (i.e., each week in the month) as the content on PULSE changes weekly. WATEC determined that 74% of

| Pageviews in the visit | Visits with this many pageviews |     | Percentage of all visits |       |
|------------------------|---------------------------------|-----|--------------------------|-------|
|                        | WATEC                           | AIU | WATEC                    | AIU   |
| 1 pageviews            | 5                               | 392 | 0.55                     | 29.95 |
| 2 pageviews            | 31                              | 107 | 3.43                     | 8.17  |
| 3 pageviews            | 52                              | 75  | 5.75                     | 5.73  |
| 4 pageviews            | 91                              | 83  | 10.07                    | 6.34  |
| 5 pageviews            | 86                              | 97  | 9.51                     | 7.41  |
| 6 pageviews            | 88                              | 69  | 9.73                     | 5.27  |
| 7 pageviews            | 61                              | 56  | 6.75                     | 4.28  |
| 8 pageviews            | 69                              | 57  | 7.63                     | 4.35  |
| 9 pageviews            | 47                              | 44  | 5.2                      | 3.36  |
| 10 pageviews           | 41                              | 45  | 4.54                     | 3.44  |
| 11 pageviews           | 57                              | 38  | 6.31                     | 2.9   |
| 12 + pageviews         | 276                             | 246 | 30.53                    | 18.79 |

(a) Visit depth

| Duration of visit | Visits with this duration |     | Percentage of all visits |       |
|-------------------|---------------------------|-----|--------------------------|-------|
|                   | WATEC                     | AIU | WATEC                    | AIU   |
| 0-10 seconds      | 30                        | 470 | 3.32                     | 35.91 |
| 11-30 seconds     | 97                        | 102 | 10.73                    | 7.79  |
| 31-60 seconds     | 102                       | 90  | 11.28                    | 6.88  |
| 61-180 seconds    | 160                       | 161 | 17.7                     | 12.3  |
| 181-600 seconds   | 184                       | 190 | 20.35                    | 14.51 |
| 601-1800 seconds  | 135                       | 178 | 14.93                    | 13.6  |
| 1801 + seconds    | 196                       | 118 | 21.68                    | 9.01  |

(b) Visit duration

| Count of visits from this visitor | Visits that were the visitor's nth visit |     | Percentage of all visits |       |
|-----------------------------------|------------------------------------------|-----|--------------------------|-------|
|                                   | WATEC                                    | AIU | WATEC                    | AIU   |
| 1 times                           | 89                                       | 554 | 9.85                     | 42.32 |
| 2 times                           | 75                                       | 243 | 8.3                      | 18.56 |
| 3 times                           | 71                                       | 146 | 7.85                     | 11.15 |
| 4 times                           | 67                                       | 94  | 7.41                     | 7.18  |
| 5 times                           | 62                                       | 68  | 6.86                     | 5.19  |
| 6 times                           | 59                                       | 40  | 6.53                     | 3.06  |
| 7 times                           | 53                                       | 33  | 5.86                     | 2.52  |
| 8 times                           | 49                                       | 24  | 5.42                     | 1.83  |
| 9 times                           | 44                                       | 19  | 4.87                     | 1.45  |
| 10 + times                        | 335                                      | 88  | 37.06                    | 6.72  |

(c) Frequency

| Previous visit was tracked | Total visits by period |     | Percentage of all visits |       |
|----------------------------|------------------------|-----|--------------------------|-------|
|                            | WATEC                  | AIU | WATEC                    | AIU   |
| First visit                | 89                     | 554 | 9.85                     | 42.32 |
| Same day                   | 291                    | 398 | 32.19                    | 30.4  |
| 1 days ago                 | 68                     | 79  | 7.52                     | 6.04  |
| 2 days ago                 | 37                     | 31  | 4.09                     | 2.37  |
| 3 days ago                 | 29                     | 21  | 3.21                     | 1.6   |
| 4 days ago                 | 27                     | 22  | 2.99                     | 1.68  |
| 5 days ago                 | 26                     | 30  | 2.88                     | 2.29  |
| 6 days ago                 | 81                     | 69  | 8.96                     | 5.27  |
| 7 days ago                 | 72                     | 58  | 7.96                     | 4.43  |
| 8-14 days ago              | 119                    | 37  | 13.16                    | 2.83  |
| 15-30 days ago             | 65                     | 10  | 7.19                     | 0.76  |

(d) Recency

Fig. 4. Simulation Results

PULSE users fulfill this constraint while only 20% of AIU's visitors revisited the site.

The interval of time since visitors last visit is called *recency* and its distribution over the recorded period of time is presented in Figure 4(d). It is expected that every visitor has to revisit the site weekly in order to consult the changes on PULSE. The results in Figure 4(d) show that within 6 days recency 91% of AIU's visitors revisited PULSE compared with

71% of WATEC's visitors. Although satisfactory, WATEC's performance is worse in this case than AIU's. The explanation lays in the fact that most PULSE users access the site mostly by using computers within laboratories. Thus, is more probable for the same IP+UA to revisit the site than an actual PULSE user.

## V. CONCLUSION AND FUTURE WORK

Test results presented in this paper show that integrated analytics are more accurate and therefore more reliable than legacy analytics. The differences between the two techniques are significant. Although preliminary tests show that cookie-based web analytics (e.g., Google Analytics) perform even worst than those presented here, future research intends to address this subject.

## 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] O. R. Zaiane, "Web usage mining for a better web-based learning," in *In Proc. of Conference on Advanced Technology for Education*, 2001, pp. 60–64.
- [4] 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.
- [5] F. Facca and P. Lanzi, "Recent developments in web usage mining research," in *Data Warehousing and Knowledge Discovery*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2003, vol. 2737, pp. 140–150.
- [6] 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.
- [7] W. A. Association, "Web analytics definitions," Web Analytics Association, 401 Edgewater Place, Suite 600, Wakefield MA 01880 USA, Draft for Public Comment, September 2008, <http://www.webanalyticsassociation.org/?page=standards>.
- [8] T. J. I. C. for Web Standards (JICWEBS), "Reporting standards. website traffic," <http://www.abc.org.uk/Corporate/AboutABC/documents/JICWEBSWebTraffic.pdf>, Auditing Bureau of Circulations electronic (ABCe), Saxon House, 211 High Street, Berkhamsted, Hertfordshire, UK, Report 1, January 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, pp. 171–190, April 2003. [Online]. Available: <http://portal.acm.org/citation.cfm?id=970781.970795>
- [10] S. Dragos and R. Dragos, "WATEC: a Web Analytics Tool for Educational Content," *KEPT 2009*, vol. Selected Papers, pp. 320–327, 2009.
- [11] —, "Web Analytics for Educational Content," *Studia Universitatis Babes-Bolyai Series Informatica*, vol. Special Issue KEPT-2009.
- [12] 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.
- [13] —, "PULSE Extended," in *The Fourth International Conference on Internet and Web Applications and Services*. Italy: IEEE Computer Society, May 2009, pp. 510–515.
- [14] —, "Current Extensions on PULSE," *Studia Universitatis Babes-Bolyai Series Informatica*, vol. LV, no. 3, pp. 51–60, 2010.