

## WATEC: A WEB ANALYTICS TOOL FOR EDUCATIONAL CONTENT

SANDA DRAGOS<sup>(1)</sup> AND RADU DRAGOS<sup>(2)</sup>

**ABSTRACT.** Web Analytics helps you to evaluate the performance of your web site. It is a series of techniques used to assess online the behavior of visitors in order to understand and optimize web usage.

There is an abundance of tools performing web analytics that can provide a baseline of statistics that indicate use levels and patterns and growth comparisons among parts of a site or over time. However, educational content needs a finer-grain evaluation up to a user-level.

This paper proposes to integrate a web analytics instrument with a learning management system for a mutual benefit. The learning management system can provide more detailed data for web analytics instrument, which eventually helps improving the learning management system to better cater to its users need.

### 1. INTRODUCTION

The Internet has become an instructional medium that has a profound impact on education by offering a new form of knowledge delivery where learner and teacher do not have to share the same physical space and time. Moreover, more educational organizations use web-based instruments as an integral part of their operations. The abundance of Learning Management Systems has resulted in an increase interest in the measurement and the evaluation of such instructional web site usage [3, 8, 13].

Most Learning Management Systems [7, 4, 16, 1, 2, 12] come as a substitute of a human tutor. The main disadvantages of such systems, as with E-learning in general, is the lack of social interaction [11]. Therefore, blended learning (also called hybrid learning) [9] which combines e-learning with traditional forms of training can be a very efficient and effective method of delivery. That is because learners can be more independent and self-reliant in their own learning, while the teacher becomes a learning facilitator.

E-learning is well suited to reconcile different learning preferences, as some students learn with pleasure and others out of plain necessity, some invest more time to gain a deep understanding, while others are merely looking for a cursory overview. In this context,

---

2000 *Mathematics Subject Classification.* 90B18, 91E45.

*Key words and phrases.* Web Usage Mining, Web Analytics, Web Metrics, e-Learning, Learning Management Systems.

improving web communication is essential to satisfy the objective of both the web site, the learning management system in our case, and its target audience, the students.

Web usage mining [15] has as strategic goals the prediction of the users behavior within the site, comparison between expected and actual web site usage, and adjustment of the web site with respect to the interests of its users. Web analytics is the part of web usage mining that focuses on improving web communication by mining and analyzing web usage data to discover interesting metrics and usage patterns [14].

There are two main technological approaches to collect data for web analytics instruments. The first method, logfile analysis, reads the logfiles in which the web server records all its transactions. The second method, page tagging, uses JavaScript on each page to notify a third-party server when a page is rendered by a web browser. Both collect data that can be processed to produce web traffic reports.

The web server log maintains a history of page requests. The World Wide Web Consortium (W3C) maintains a standard format [10] for web server log files. They contain information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent and referrer. These files are usually allowed only to the webmaster or other privileged person.

## 2. OUR PROPOSAL

The idea of this new line of research started from two existing instruments: a learning management system, called PULSE [5, 6] and a web analytics instrument. We will call the latter WATEC, which stands for Web Analytics Tool for Educational Content. The advantage of integrating the two is twofold:

- **Improve PULSE.** It is helpful to have an informed learning management system that continually "educates" itself about the requirements of its users as a result of the feedback offered by the web analytics tool. The usage information has the potential to be used by WATEC to evaluate the effectiveness of PULSE. For instance it can determine whether:
  - (1) students are finding the essential web pages;
  - (2) students follow the optimal paths in reaching the sought information;
  - (3) students are spending (an how much) time on specific web-pages;
  - (4) a web pages is unnecessary;
  - (5) all pages are tailored for the used browsers or resolutions;
  - (6) there are changes in online traffic patterns and behavior over time;
- **Help students.** Share from navigational experience of other students on what they found useful. Learning is often aided with the inclusion of other kinds of data such as the structure of the learning management system or its usage information (e.g. most visited pages or most used paths in reaching the sought information).

Moreover, each of these two instruments can be tailored to improve the performance of the other. For instance, PULSE can feed more specific information to be logged which

will result in a better interpretation for WATEC, while feedback offered by the latter helps understand the strong points and the shortcomings and improvements of PULSE.

The most important attribute of PULSE that qualifies it for the study of individual student behavior patterns is the authentication of its users (students, tutors or its administrator). Logging such information can result in statistics such as the top of ten authenticated users, the average of the time (on daily, weekly or monthly bases) spent on a site by a single user or top (or number) of pages visited by a specific user.

The main advantages of WATEC over other such instruments can be distinguished into three classes:

- (1) It is simple and focuses on our main objective that is to constantly monitor PULSE and adapt it to the needs of their users. Thus it can always be tailored to accommodate more desired features.
- (2) It records every transaction PULSE makes. Page tagging may not be able to record all transactions as it relies on the visitors' browsers cooperation. For instance JavaScript could be disabled.
- (3) Some useful analysis can be made accessible for students.

### 3. CURRENT DEVELOPMENT

WATEC is a PHP web statistics tool that gathers site-usage information into a MySQL database and creates analysis such as yearly, monthly, weekly, daily and hourly statistics on the number of hits, the number of visits, the number of pages and the number of sites. All these results are presented graphically and in table format. Such overall statistics over three years are depicted in Figure 1.

WATEC allows filtering all logged data on different specific criteria as presented in Figure 2. The filtering can be done either by exact match or by substrings.

All data presented in Figure 1 is filtered to refer only to PULSE sites. It can be observed that the activity on PULSE intensified during the semester and specifically around laboratory sessions when students had to login with PULSE to get their assignments and to consult lectures notes or other theoretical support that PULSE has to offer. Thus there are some distinct peaks at a distance of 7 days of each other. The final peak is around the day of the final examination. The last half of April was the Easter Holiday and the activity dropped, as it did after the examination.

Another important aspect that can be deduced from these graphs is that the number of visits (distinct IP addresses) and sites (distinct hostIP addresses) is very low compared with the number of hits. That is because laboratory sessions are held in the same classrooms and therefore, students use the same computers that are assigned with the same IPs. In this case one IP represents more visitors, except the situations when a student is using his/her personal computer from his/her home.

WATEC also generates different lists ranked by the number of hits assigned to them. They are collections of:

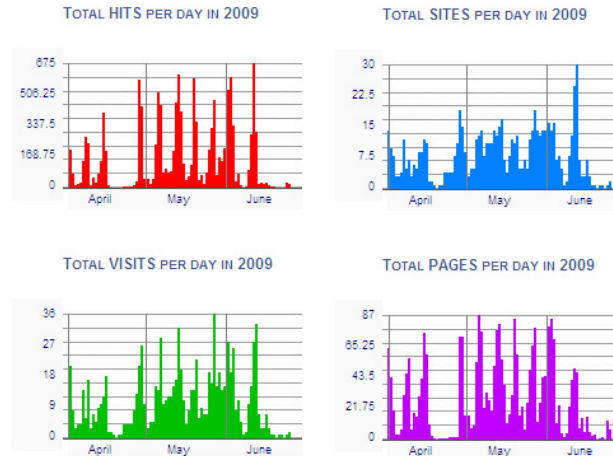


FIGURE 1. Generated Traffic

SELECT ONE OF THE CONDITIONS:

|                                     |             |     |     |                         |                                     |                                       |                                      |                           |                                       |
|-------------------------------------|-------------|-----|-----|-------------------------|-------------------------------------|---------------------------------------|--------------------------------------|---------------------------|---------------------------------------|
| <input checked="" type="checkbox"/> | DATE:       | Jul | 11  | 2009                    | <input type="radio"/> Day precision | <input type="radio"/> Month precision | <input type="radio"/> Year precision | <input type="radio"/> all | <input checked="" type="radio"/> none |
| <input type="checkbox"/>            | IP:         | is  | --- | select 'IP' ---         | or contains                         |                                       |                                      |                           |                                       |
| <input type="checkbox"/>            | HOSTIP:     | is  | --- | select 'hostIP' ---     | or contains                         |                                       |                                      |                           |                                       |
| <input type="checkbox"/>            | HOSTNAME:   | is  | --- | select 'hostName' ---   | or contains                         |                                       |                                      |                           |                                       |
| <input type="checkbox"/>            | USERAGENT:  | is  | --- | select 'userAgent' ---  | or contains                         |                                       |                                      |                           |                                       |
| <input checked="" type="checkbox"/> | ACCFILE:    | is  | --- | select 'accfile' ---    | or contains                         | PULSE                                 |                                      |                           |                                       |
| <input type="checkbox"/>            | RESOLUTION: | is  | --- | select 'resolution' --- | or contains                         |                                       |                                      |                           |                                       |
| <input type="checkbox"/>            | RATIO:      | is  | --- | select 'ratio' ---      | or contains                         |                                       |                                      |                           |                                       |
| <input type="checkbox"/>            | REFERER:    | is  | --- | select 'referer' ---    | or contains                         |                                       |                                      |                           |                                       |

FIGURE 2. Filtering logged information

**hostnames that accessed the site:** A list is provided with all hostnames corresponding to the host IP addresses that accessed the site. The list is ordered in a decreasing order by the number of hits from each hostname. There are also provided three more lists with Top Level Domains (i.e., edu, net, com), second level domains (i.e., ubbcluj.ro, googlebot.com, rdsnet.ro), and third level domains (i.e., search.msn.com, crawl.yahoo.net, staff.ubbcluj.ro). These statistics offer us a view on who is visiting the site in terms of geographic location

(e.g., ro, ie, de, it) or the searching engines they used (e.g., googlebot.com, search.msn.com, crawl.yahoo.net).

| Top 41 sites |      |                                    |    | Internet domains              |      |                 |    |
|--------------|------|------------------------------------|----|-------------------------------|------|-----------------|----|
| #            | hits | Hotname                            | %  | #                             | hits | Hotname         | %  |
| 1            | 3137 | 230.9.122.86.static.cluj.rdsnet.ro | 24 | 1                             | 9329 | ro              | 70 |
| 2            | 3048 | campus.cs.ubbcluj.ro               | 23 | 2                             | 3770 | UNDEFINED       | 28 |
| 3            | 1734 | 193.231.20.20                      | 13 | 3                             | 155  | com             | 1  |
| 4            | 1218 | apollo1.ubbcluj.ro                 | 9  | 4                             | 55   | net             | 0  |
| 5            | 1148 | 78.96.153.97                       | 9  | 2 <sup>nd</sup> Level Domains |      |                 |    |
| 6            | 718  | Marasti-IF.east.utcluj.ro          | 5  | 1                             | 4875 | ubbcluj.ro      | 37 |
| 7            | 474  | 194.102.61.170                     | 4  | 2                             | 3770 | UNDEFINED       | 28 |
| 8            | 314  | li.cj.edu.ro                       | 2  | 3                             | 3139 | rdsnet.ro       | 24 |
| 9            | 250  | apollo14.UBBCluj.Ro                | 2  | 4                             | 904  | utcluj.ro       | 7  |
| 10           | 242  | 89.137.7.219                       | 2  | 5                             | 314  | edu.ro          | 2  |
| 11           | 194  | apollo16.ubbcluj.ro                | 1  | 6                             | 132  | intellisync.com | 1  |
| 12           | 186  | marasti-iib2.east.utcluj.ro        | 1  | 7                             | 54   | idilis.net      | 0  |
| 13           | 132  | nat.intellisync.com                | 1  | 8                             | 46   | bizarnet.ro     | 0  |
| 14           | 72   | 193.226.40.56                      | 1  | 9                             | 27   | astral.ro       | 0  |
| 15           | 66   | 1306_10.scs.ubbcluj.ro             | 0  | 10                            | 22   | it4web.ro       | 0  |
| 16           | 54   | btsec.cj.idilis.net                | 0  | 11                            | 15   | searchme.com    | 0  |
| 17           | 46   | proxy2.bizarnet.ro                 | 0  | 12                            | 5    | googlebot.com   | 0  |
| 18           | 45   | 1306_07.scs.ubbcluj.ro             | 0  | 13                            | 2    | msn.com         | 0  |
| 19           | 44   | 89.238.209.66                      | 0  | 14                            | 1    | yahoo.net       | 0  |

FIGURE 3. Hostnames that accessed the site

Figure 3 presents the list of hostnames from where PULSE was accessed within the 3 months (April, May, June) of 2009. As it can be seen on the list of Internet domains there are only three: .ro because all our students are from Romania, .com and .net from the searching engines or other internet providers (e.g. idilis.net) as depicted in the list of the second level domains.

**pages that were accessed:** The list of pages offers us a view on the most "interesting" pages on the site.

| Top 435 pages |      |                                                  |    | Top of directories |       |                       |     |
|---------------|------|--------------------------------------------------|----|--------------------|-------|-----------------------|-----|
| #             | hits | URL                                              | %  | #                  | hits  | Directory             | %   |
| 1             | 4078 | /~sanda/teaching/PULSE.php?2008-2009/SO1         | 26 | 1                  | 15374 | /~sanda/teaching      | 100 |
| 2             | 2121 | /~sanda/teaching/PULSE.php                       | 14 | 2                  | 17    | /~sanda/_new/teaching | 0   |
| 3             | 1107 | /~sanda/teaching/PULSE.php?2008-2009/SO1&cursori | 7  | 3                  | 16    | /~sanda/publications  | 0   |
| 4             | 398  | /~sanda/teaching/PULSE.php?2008-2009/SO1&logout  | 3  |                    |       |                       |     |
| 5             | 349  | /~sanda/teaching/PULSE.php?2008-2009/SO1&lab8    | 2  |                    |       |                       |     |

FIGURE 4. Pages that were accessed

Figure 4 lists the pages with most hits. This statistics make sense as the only subject that used PULSE in this last semester was Operating Systems. Thus, the first entry in the list contains the substring "SO". There were also only three directories to contain PULSE related files and they are also listed in Figure 4.

**user agents used:** The most important statistics collected from user agents are the operating systems used by the visitors and their browser type and version. A sample of such lists is presented in Figure 5.

| # | hits  | Operating System | %  | #  | hits | Browser     | Type | %  |
|---|-------|------------------|----|----|------|-------------|------|----|
| 1 | 13775 | Windows XP       | 89 | 1  | 6251 | IE 7.0      | Ⓜ    | 41 |
| 2 | 1185  | Windows Vista    | 8  | 2  | 4925 | Firefox 1.9 | Ⓜ    | 32 |
| 3 | 228   | UNKNOWN          | 1  | 3  | 1853 | IE 6.0      | Ⓜ    | 12 |
| 4 | 207   | Linux ubuntu     | 1  | 4  | 1537 | Firefox 1.8 | Ⓜ    | 10 |
| 5 | 12    | Linux            | 0  | 5  | 414  | OP 9.63     | Ⓜ    | 3  |
|   |       |                  |    | 6  | 153  | OP 9.64     | Ⓜ    | 1  |
|   |       |                  |    | 7  | 111  | IE 8.0      | Ⓜ    | 1  |
|   |       |                  |    | 8  | 88   | OP 9.52     | Ⓜ    | 1  |
|   |       |                  |    | 9  | 20   | Saf         | Ⓜ    | 0  |
|   |       |                  |    | 10 | 18   | UNKNOWN     | ?    | 0  |
|   |       |                  |    | 11 | 15   | Google      | ?    | 0  |
|   |       |                  |    | 12 | 11   | OP 9.23     | Ⓜ    | 0  |
|   |       |                  |    | 13 | 5    | Msn         | ?    | 0  |
|   |       |                  |    | 14 | 3    | Firefox 1.7 | Ⓜ    | 0  |
|   |       |                  |    | 15 | 3    | Yahoo       | ?    | 0  |

FIGURE 5. User agents used

The operating systems list can indicate the device used to access the site. For instance "Windows CE" is used by minimalistic computers and embedded systems such as personal digital assistants (PDAs) or mobile phones. User agents can also indicate if a visitor arrived at the site through search engines. As presented in Figure 5 the largest amount of such traffic is coming from search engines such as Google, Yahoo, and MSN.

**referrers:** Is another way of determining where people are visiting from, as a referrer is the URL of a previous item which led to this request. Referrers can be local pages (a page within the site) or an external page (which can again be a search engine). Statistics on local referrers can help content site optimization by determining which content areas have the most affinity.

**screen resolutions:** As the operating system, screen resolution can indicate the device used. The list of resolutions used in the period of time aforementioned by PULSE users is presented in Figure 6.

#### 4. FUTURE WORK

Out of the three fundamental questions (Who?, What? and Why?) WATEC answers the first two of them. It determines who is visiting PULSE and what they are looking for. The most important question remains 'Why?'. We are working to extend WATEC to answer this last question, more specifically: Why doesn't one student ever visit the site?

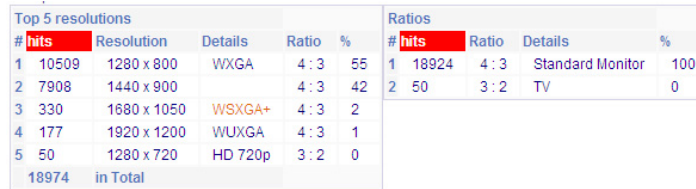


FIGURE 6. Screen resolutions used

Why does another student visit it twice a day? Is the material helpful? What material is the most helpful?

From the current statistics it results that PULSE is helpful as it recorded around 3000 hits within last month. However, our goal is to obtain more meaningful statistics by implementing the two following strategies:

**Visitor segmentation:** Segmentation isolates the behavior of certain types of on-line visitors. By using PULSE's log-in phase, individual student's site accesses can be located within collected data. Thus, segmentation can be performed based on demographics such as gender, year of study, line of study, marks.

Despite the fact that each person's learning requirements may be different, there are often wide areas of overlap between individuals that can be mutually beneficial. Similarity in learning needs defining functional communities of learners.

**Testing and experimentation:** By using slight variations, it is possible to determine which minor differences make the biggest difference. The same content presented in different format (e.g., text versus graphical/multimedia, pdf versus presentation) can have a different impact on students/visitors.

## 5. CONCLUSION

Web analytics instruments can be used as a starting point in understanding student behavior and learning patterns while using an e-learning instrument. However, the analytical results offered by such tools are too coarse and may not be able to give an insight on individual student behavior. Tiding the logged information to the structure of the specific learning management system may increase the efficiency of the web analytics tool.

We proposed in this paper the integration of two existing instruments: PULSE an learning management system used to support a professor to provide high quality education for a large number of students in an field of rapid changes and practical aspects, and WATEC a web analytics tool used to constantly monitor web sites and evaluate their adaptiveness to the need of their users. The benefit of such an fusion is mutual as PULSE can be easily modified to provide essential data to WATEC, while the latter interprets such information in order to provide a better understanding users needs. WATEC can also be used to reveal PULSE's shortcoming from this perspective so that they can be addressed.

## REFERENCES

- [1] *Apex Learning*, July 2009. <http://www.apexlearning.com/>.
- [2] *Blackboard*, July 2009. <http://www.blackboard.com/>.
- [3] P. DESIKAN, C. DELONG, K. BEEMANAPALLI, A. BOSE, AND J. SRIVASTAVA, *Data Mining in E-Learning*, WIT Press, 2006, ch. Web Mining for Self-Directed E-learning, pp. 21–40.
- [4] DOKEOS, *Dokeos Open Source e-Learning*, July 2009. <http://www.dokeos.com/>.
- [5] S. DRAGOS, *PULSE - a PHP Utility used in Laboratories for Student Evaluation*, in International Conference on Informatics Education Europe II (IEEEII), Thessaloniki, Greece, Nov. 2007, pp. 306–314.
- [6] ———, *Pulse extended*, in Fourth International Conference on Internet and Web Applications and Services (ICIW), M. Perry, H. Sasaki, M. Ehmann, G. O. Bellot, and O. Dini, eds., Venice/Mestre, Italy, May 2009, IEEE Computer Society, pp. 510–515.
- [7] ELEARNING INDIA, *Learning Management Systems - LMS*, July 2009. <http://elearning-india.com/content/blogcategory/19/38/>.
- [8] K. FANSLER AND R. RIEGLE, *A model of online instructional design analytics*, in 20th Annual Conference on Distance Learning and Learning, 2004.
- [9] R. GARRISON AND H. KANUKA, *Blended learning: Uncovering its transformative potential in higher education*, The Internet and Higher Education, 7 (2004), pp. 95–105.
- [10] P. M. HALLAM-BAKER AND B. BEHLENDORF, *Extended log file format*, Working Draft WD-logfile-960323, World Wide Web Consortium (W3C).
- [11] A. HEINZE AND C. PROCTER, *Reflections on the use of blended learning*, in Education in a Changing Environment, University of Salford, 2004. ISBN: 0902896806.
- [12] IMC AG, *eLearning Suite CLIX*, July 2009. <http://www.im-c.de/Products/eLearning-Suite/>.
- [13] D. MONK, *Using data mining for e-learning decision making*, The Electronic Journal of e-Learning, 3 (2005), pp. 41–54.
- [14] J.-P. NORGUET, E. ZIMNYI, AND R. STEINBERGER, *SOFSEM 2006: Theory and Practice of Computer Science*, vol. 3831, Springer Berlin / Heidelberg, 2006, ch. Improving Web Sites with Web Usage Mining, Web Content Mining, and Semantic Analysis, pp. 430–439. ISBN 978-3-540-31198-0.
- [15] J. SRIVASTAVA, R. COOLEY, M. DESHPANDE, AND P.-N. TAN, *Web usage mining: Discovery and applications of usage patterns from web data*, SIGKDD Explorations, 1 (2000), pp. 12–23.
- [16] UNIVERSITY OF ZURICH IN ASSOCIATION WITH THE COMMUNITY, *OLAT - Open Source LMS*, July 2009. <http://www.olat.org/website/en/html/>.

<sup>(1)</sup> BABES-BOLYAI UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE  
*E-mail address:* sanda@cs.ubbcluj.ro

<sup>(2)</sup> BABES-BOLYAI UNIVERSITY, COMMUNICATION CENTER  
*E-mail address:* radu.dragos@ubbcluj.ro