Behavioral Patterns of Users Accessing Learning Resources in a Controlled Environment

Antonio Maratea, Alfredo Petrosino, Mario Manzo

Abstract: *learning resources are more than ever published in e-learning platforms and made remotely accessible to the public, and more than ever it is essential to have tools that measure feedback on the resource effectiveness, assimilation time and user satisfaction. Based on log data from a group of graduated students following a specialized course, in this paper it is shown that access to resources follows a very similar and general pattern. Three main behavioral stages can be recognized in the clicking distribution, called orientation, evaluation and assimilation. The clicking distribution is then used to derive session time, total time of fruition and number of sessions.*

Key words: *log analysis, session estimation, clicking distribution.*

INTRODUCTION

In many structured learning scenarios, monitoring user behavior and total time spent on resources is an essential feedback on the effectiveness of the resource itself and an implicit measure of user satisfaction. Log data and user click modeling have been deeply investigated during recent years, mainly in areas of internet marketing, banner effectiveness evaluation, website navigation, implicit feedback and personalized recommendations [1,2,3,5], but the case of BS or MS graduated students from a University accessing learning resources is different from a generic user surfing a web page: the learning environment is without commercial distractions or out of scope content and it is reasonable to assume a more focused interest, a more stable concentration ability and a more effective strategy of study. At the same time, estimating the time spent on the resource is not an easy task, due to log data commonly tracking only clicks and not actual starts and ends of sessions.

In this paper a model for user clicking behavior in a controlled environment and a method based on it to estimate the session time, the number of sessions and the total time of fruition of a resource are proposed.

LOG ANALYSIS

The description and analysis of user behavior is based on logs of student accessing an e-learning platform with a predefined task.

Data

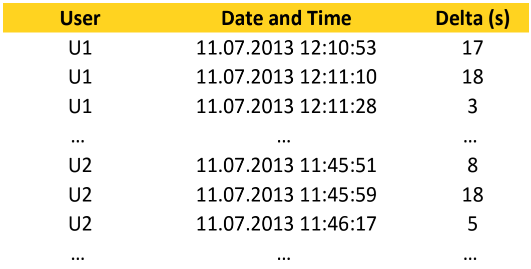
Data were collected from July 2013 to February 2014 and concern 16 graduate users that had to complete a mandatory program within a publicly financed project.

The program consisted in 66 lessons, each one during 27 minutes and 8 seconds, that were embedded in SCORM compliant packages (to keep track easily of the progress of each lesson) and that each student had to complete from home before a given date (28/02/2014).

Logs of user clicks were obtained from the Moodle e-learning platform that embedded the SCORM packages, hence no navigation of any single resource is tracked: clicks show how each users behaves in accessing resources to be learnt from the platform and are used to estimate how much time he or she actually spent on it.

Preprocessing

Logs have been anonymized, converted into intervals (expressed in seconds) and analyzed. Only the date and time of the clicks and the anonymized user ID have been kept – no auxiliary variables were collected. See Figure 1 to see an example.



User	Date and Time	Delta (s)
U1	11.07.2013 12:10:53	17
U1	11.07.2013 12:11:10	18
U1	11.07.2013 12:11:28	3
...	...	...
U2	11.07.2013 11:45:51	8
U2	11.07.2013 11:45:59	18
U2	11.07.2013 11:46:17	5
...	...	...

Figure 1: structure of the log file

Looking to the data for each user, it was clear that some intervals of time abnormally high – exceeding several hours from one click to the next – were due to the switch from a day to the next one. After a careful check, all the values due to the day switch were removed.

Another observation was that the large majority of click were within one second from the previous one, likely due to the first steps in finding the resource on the platform: also these clicks were considered uninformative and removed.

To obtain meaningful frequency histograms, data were binned with a log scale, to account for the spread of clicks becoming wider as time goes by. The effect of the log transformation on the frequency histogram of click intervals for a single user can be seen in Figure 2, where 1-second clicks are not yet removed.

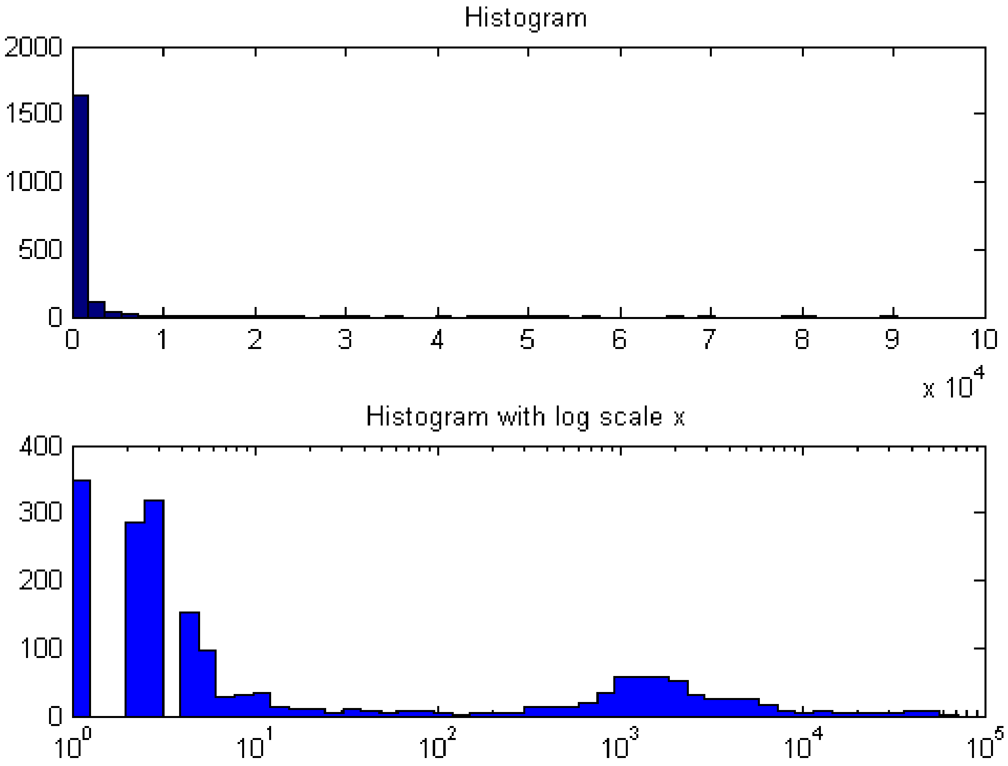


Figure 2: frequency histograms on linear (top) and logarithmic scale (bottom).
Y axis represent the absolute frequency of clicks
and X axis the interval time between clicks in seconds.

Fitting distributions

The frequency histograms of the various users were surprisingly similar (see Figure 6), and the average frequency distribution for the clicks, shown in Figure 3, well represents a common behavioral pattern. As it can be seen, the distribution is neatly divided in two concentrated areas, with almost no activity in between.

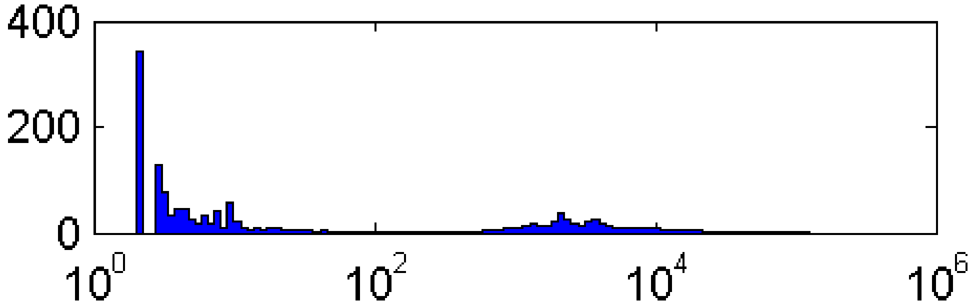


Figure 3: average frequency histogram among all users, H1.
Y axis represent the absolute frequency of clicks
and X axis the interval time between clicks in seconds.

First an Otsu thresholding [6] was applied to split the histogram in two halves with the maximum inter-class variability

$$\sigma^2_\omega(t)= \omega_0(t)\sigma^2_0(t)+ \omega_1(t)\sigma^2_1(t) \qquad (1)$$

weights $\omega_{0,1}$ are the probabilities of the two classes being separated by a threshold t and $\sigma^2_{0,1}$ are variances of these two classes. The original algorithm is conceived for 256 gray levels, so it has been extended to provide for more levels.

Then a Kolmogorov-Smirnov test [4] for goodness of fit was performed, testing all major probability distributions to find the best fit on the two halves.

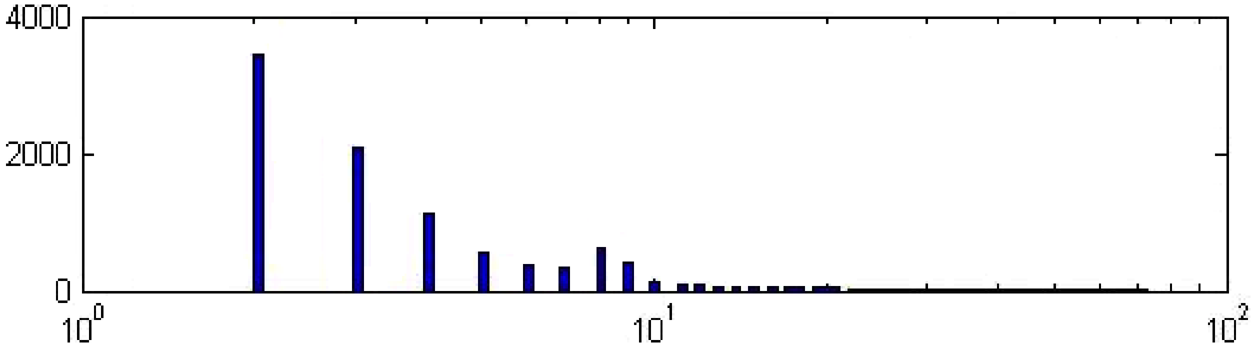


Figure 4: the frequency histogram corresponding to the first half
of the average histogram among all users.

While the shape reminds a Beta, the histogram corresponding to the first half on the left, called *H1L*, did not fit with any of the major probability distribution checked (generalized extreme value, inversegaussian, generalized pareto, birnbaumsaunders, lognormal, loglogistic, gamma, nakagami, rayleigh, weibull, rician, tlocationscale, logistic, normal, exponential, extreme value), being extremely skewed.

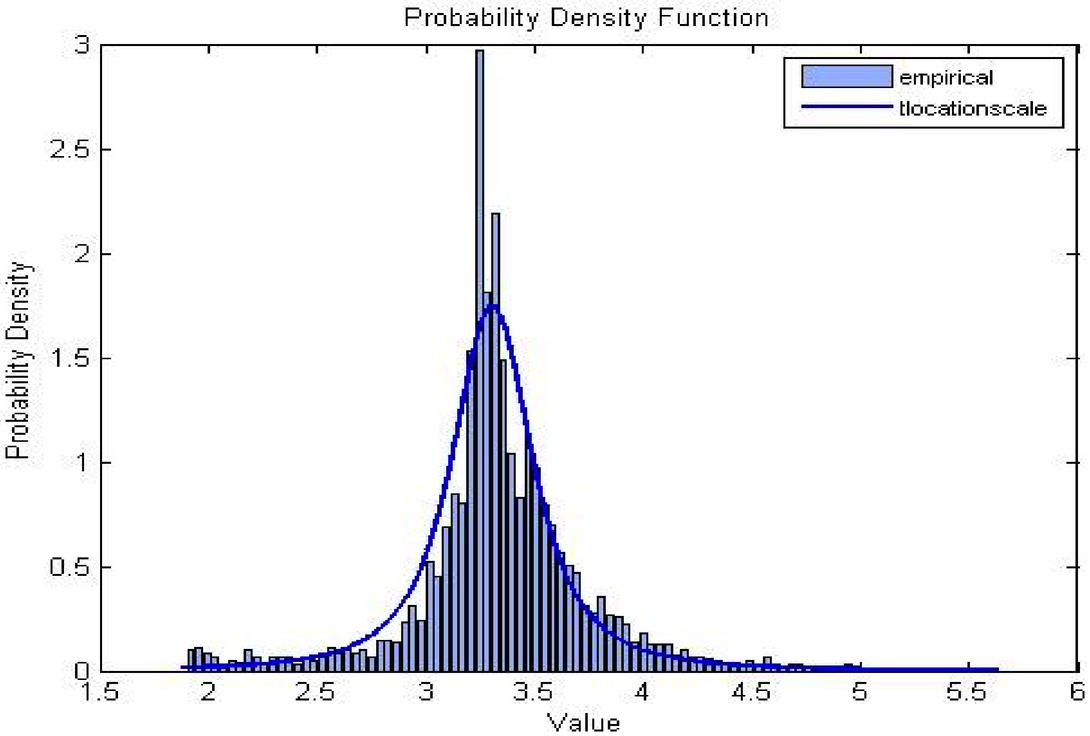


Figure 5: fitting distribution overlapped to the the frequency histogram corresponding to the second half of the average histogram among all users.

On the other hand, the second half, called *H1R*, is nicely fitted by a T distribution, with location value of 3.3052 and scale values of 0.2011. Please note that the T distribution has an heavier tail with respect to a Gaussian.

Estimating session time, total time and the number of sessions

The average session time *ST* for all users was obtained subtracting the median of the distribution in H1L, called S_{low}, from the median of the distribution in H1R, called S_{high}.

$$ST = S_{high} - S_{low} \qquad (2)$$

Called t_{ui} the Delta values in seconds as recorded in the log file (see Figure 1) for each user *u*, the total fruition time for each user T_u was estimated summing all the time intervals of that user within the session, that is summing for each user all t_i greater than S_{low} and smaller than S_{high}.

$$T_u = SUM(t_{ui}) \qquad s.t. \; S_{low} < t_{ui} < S_{high} \qquad (3)$$

Having the total time and the average session time it is possible to estimate the number of sessions ns_u for each user in the following way:

$$ns_u = ceil(T_u/ST) \qquad (4)$$

where *ceil()* is the next integer approximation function.

RESULTS AND DISCUSSION

The average session time, the average number of sessions and the average total time on the platform for each user are 90.6 minutes, 156 sessions and 234. 4 hours respectively. From a comparison of the total lesson time (29 hours and 50 minutes) to the total time spent on the platform, considering that also tests and exercises were given to users in order to evaluate their learning progress, the ratio between the lesson time and time actually spent on the platform to learn that lesson is circa 1:8.

The shape of the average distribution of clicks shown in Figure 3 and the fitting probability distributions shown thereafter allow several interesting considerations:

First, in a controlled and structured learning scenario with experienced students and predefined mandatory tasks, even if the time of fruition is freely chosen, access to resources follows a very similar pattern among the users.

Second, from the three main modes of the distribution three main behavioral stages with soft transitions can be recognized, here called *orientation*, *evaluation* and *assimilation*. The orientation is the first stage where in the first few seconds of navigation the number of clicks is very high due to the quest for the wanted resource and the need for orientation in the platform. In this stage, most of the pages are abandoned and most of the resources excluded within 1 to 5 seconds. Then there is the evaluation stage, were the reached resources are considered as interesting candidates to be studied and receive attention for 8 to 10 seconds for a deeper check, so to take a decision. Finally there is the assimilation stage, where within 30 seconds candidates are finally chosen and the learning activity actually starts.

Third, attention and concentration measured by session time have a very similar lasting and a symmetric unimodal distribution. This can be due to the homogenous population or to biological constraints.

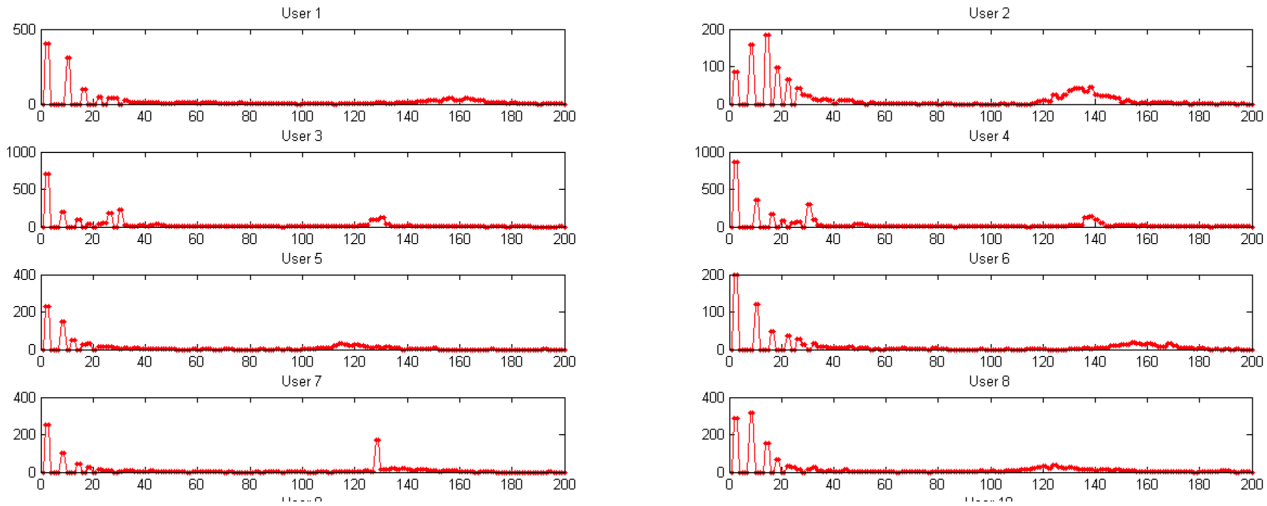


Figure 6: click frequency for 8 sample users. Y axis represent the absolute frequency, while X axis represents the number of the histogram bin.

CONCLUSIONS AND FUTURE WORK

It has been shown that remote access to learning resources from a homogenous group – even if free from time schedules – follows a very similar pattern. Three main behavioral stages in the clicking distribution of users have been recognized, called orientation, evaluation and assimilation. The clicking distribution has been finally used to derive session time, total time of fruition, average number of sessions for user and to estimate the multiplication factor between lesson time and assimilation time. Further studies are necessary to confirm the generality of the presented results.

REFERENCES

[1] Ferone A., Manzo M., Maratea A., Petrosino A. Tracking e-learning user sessions, IX National Conference of eLearnig Italian Society, 2013, 139-142

[2] JOACHIMS, Thorsten, et al. Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Acm, 2005. p. 154-161.

[3] LIU, Jiahui; DOLAN, Peter; PEDERSEN, Elin Rønby. Personalized news recommendation based on click behavior. In: Proceedings of the 15th international conference on Intelligent user interfaces. ACM, 2010. p. 31-40.

[4] Massey Jr, Frank J. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 1951, 46.253: 68-78.

[5] MURPHY, Jamie. Surfers and Searchers An Examination of Web-site Visitors' Clicking Behavior. Cornell Hotel and Restaurant Administration Quarterly, 1999, 40.2: 84-95.

[6] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber., 1979, 9 (1): 62–66.

ABOUT THE AUTHORS

Antonio Maratea, University of Naples "Parthenope", Department of Science and Technologies, Centro Direzionale di Napoli, Isola C4, 80143 Napoli, Italy, e-mail: antonio.maratea@uniparthenope.it

Alfredo Petrosino, University of Naples "Parthenope", Department of Science and Technologies, Centro Direzionale di Napoli, Isola C4, 80143 Napoli, Italy, e-mail: alfredo.petrosino@uniparthenope.it

Mario Manzo, University of Naples "Parthenope", Center of Information Technology Services, Amm. F. Acton street, 38, 80133 Napoli, Italy, e-mail: mario.manzo@uniparthenope.it

The paper has been reviewed.