Leveraging learners' activity logs for course reading analytics using session-based indicators

Madjid Sadallah*

Computer Science Department,
University of Bejaia,
06000 Bejaia, Algeria
and
CERIST Research Centre,
Algiers 16030, Algeria
Email: msadallah@cerist.dz
*Corresponding author

Benoît Encelle

University of Lyon 1,
LIRIS – UMR 5205 CNRS,
Lyon, France
Email: bencelle@liris.cnrs.fr

Azze-Eddine Maredj

CERIST Research Centre,
Algiers 16030, Algeria
Email: amaredj@cerist.dz

Yannick Prié

University of Nantes,
LINA – UMR 6241 CNRS,
Lyon, France
Email: yannick.prie@univ-nantes.fr

Abstract: A challenge that course authors face when reviewing their contents is to detect how to improve their courses in order to meet the expectations of their learners. In this paper, we propose an analytical approach that exploits learners' logs of reading to provide authors with insightful data about the consumption of their courses. We first model reading activity using the concept of reading-session and propose a new and efficient session identification. We then elaborate a list of indicators computed using learners' reading sessions that allow to represent their behaviour and to infer their needs. We evaluate our proposals with course authors and learners using logs from a major e-learning platform. Interesting results were found. This demonstrates the effectiveness of the approach in identifying aspects and parts of a course that may prevent it from being easily read and understood, and for guiding the authors through the analysis and review tasks.

Biographical notes: Madjid Sadallah is a PhD Student at the University of Bejaia and a Research Assistant at the Algerian Research Centre on Scientific and Technical Information CERIST. He worked on structured documents and multimedia aspects such as authorship, visualisation, adaptation and distribution. His current research focuses on knowledge discovery and visual analytics in the domains of education and knowledge management.

Benoît Encelle holds an Engineer's degree and a PhD in Computer Science. He currently works as an Assistant Professor at the University of Lyon, LIRIS CNRS lab. His research interests are mainly focused on interactive adaptive systems, assistive technologies and human computer interaction, trying to improve systems accessibility, usability and utility. He has a strong interest in assistance and document (re-)engineering in trace-based systems.

Azze-Eddine Maredj acquired his Doctorate in Computer Science at the University of Science and Technology Houari Boumediane (Algeria). He is a Research Scientist at the Algerian Research Centre on Scientific and Technical Information CERIST where he Leads the "*Multimedia Systems and Documents*" team. His Research interests are related to the development of systems for the authoring and the retrieval of structured and multimedia documents.

Yannick Prié is a Professor of Computer Science and Interdisciplinarity at the University of Nantes. His research interests are related to activity analysis, human-computer interaction, document and knowledge engineering, and visual analytics. He has built video annotation systems, and reflective trace-based systems within several application domains such as technology-enhanced learning, knowledge work, science and technology studies, etc.

This paper is a revised and expanded version of a paper entitled 'Towards Reading Session-Based Indicators in Educational Reading Analytics' presented at the '10th European Conference on Technology Enhanced Learning EC-TEL 2015', Toledo, Spain, 15–18 September 2015.

1   Introduction

The global adoption of information technologies in education has allowed the popularisation of distance education, the emergence of online educational portals and an increase in online course enrolment (Jo et al., 2014). As a consequence, more and more educational resources are being made available online by thousands of authors, and are being accessed by millions of learners each day, be it for rapid consultation or thorough and active reading. The main objective of any course author is to facilitate the appropriation of the knowledge he wishes to convey in his document by offering

consistent content while being easy to read and understand. However, authoring and maintaining digital learning content is not an easy task. This is partly due to the intrinsic difficulty of structuring ideas and writing, all the more so when the rapid evolution of the available means of presenting information (new modalities, interactivity) is not fully mastered, producing more readers' misperceptions, misinterpretations and misunderstandings than with classical textbook writing. Another reason is that the readership is mostly not known at the time of writing: the great heterogeneity in learners' profiles (age, background, need, etc.) and online behaviours (pace of learning: time/space/frequency preferences, usages, etc.) induces a wide diversity of reading usages and interaction possibilities.

Fortunately, the educational settings are cyclic, each arrival of new students gives time to course authors to adapt and improve their learning material. Subsequently, the contents evolve and hopefully get better (e.g., more precise, more comprehensible, more adapted to their students need, etc.). The popularisation of educational platforms with logging capabilities promotes the use of automated approaches to analyse learners' logs in order to discover previously inaccessible knowledge. Consequently, the authors can get automated feedback on the learning experience. This can help them infer what is actually learned from the courses, and make informed decisions on how their courses should be improved, from local clarifications to deeper restructuration. Revising contents by considering readers' perspective is a successful strategy to guarantee better level of understanding (Cho and MacArthur, 2011). However, little attention has been devoted to assess what learners actually understand (Dascalu et al., 2014). This is in part due to the difficulty to monitor comprehension, which requires sensitive observation skills and an active learning environment (Pattanasri et al., 2012).

In this paper, we propose a learning analysis approach that exploits learners' reading logs in order to provide authors with insights that would help them maintaining and evolving their online courses. In this perspective, we first model reading activity using the concept of "*reading session*" for denoting learner's active reading periods. We then propose several reading indicators that we compute using the resulted reading sessions. We analyse the values of these indicators within a course to spot possible reading and comprehension issues that learners may face. We illustrate these proposals on the consumption logs of several courses provided by a major French e-learning platform. We evaluate the quality of our reading session detection algorithm by comparing the results with other methods, and estimate the usefulness of our indicators through a survey for authors of courses.

This article is a revised and expanded version of the work originally presented in Sadallah et al. (2015). The remainder of this paper is structured as follows. Section 2 presents the related research concerning monitoring e-learning, reading activity and using session-based indicators in education. Section 3 introduces a new algorithm for computing learners' sessions using logs of reading activity, and details a set of activity indicators constructed from learners' reading sessions. Section 4 presents the evaluation methodology used for answering three research questions: (1) the capabilities of the session identification algorithm; (2) the relevance of the proposed indicators; and (3) the effectiveness of the detected reading issues. Section 5 discusses our results before concluding.

2   Related research

2.1   Monitoring e-learning

Monitoring learning is a common practice that helps instructors regularly evaluate the effectiveness of their teaching to make subsequent instructional decisions and interventions. It is usually achieved using data that describe the learning experience, often reported explicitly through direct interaction with learners (e.g., face-to-face observations, self-report questionnaires and interviews) (Brown et al., 2016). In distant and online learning, these forms of interaction are most of the time intrusive and expensive, and do not fit well with the heterogeneity of learner profiles, nor do they quickly reflect changes in the learning settings (Thomas, 2014). The emergence of e-learning platforms has led to the development of more unobtrusive, objective and reliable methods for gathering data, using the logging capabilities afforded by these platforms (Cocea and Weibelzahl, 2011). Different data sources can be instrumented and monitored, including learners' clickstreams (Siemens, 2013), eyes movements (Copeland et al., 2015), participation (Xing et al., 2015), and/or assessments (Fidalgo-Blanco et al., 2015; Snodgrass-Rangel et al., 2015). During the data analysis process, the captured data undergo different transformations, in order to be finally translated into understandable and exploitable human knowledge. This process is largely covered by the field of *Learning analytics* which aims to develop and promote methods of "measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Siemens, 2012). Different analytics and data mining methods can be applied on the tracked data (e.g., clustering, classification, association rule mining, etc.) for different purposes. For instance, they are used to study learners' retention and learning outcomes (Edwards et al., 2017), learners' engagement and motivation (D'Mello et al., 2017; Tempelaar et al., 2017), self-regulated learning (Kizilcec et al., 2017), learning strategies (Kovanovic et al., 2015), and modelling learners' misconceptions (Liu et al., 2016).

2.2   Monitoring and analysing reading activity

Online reading is a fundamental activity in e-learning; it consists on a series of page access actions performed on the learning platform. By recording and studying these actions, it becomes possible to reconstruct then examine the actual reading experience of learners. It becomes even possible to predict with good accuracy the future usages and behaviours (Hauger et al., 2011; Zahoor et al., 2015). This usually requires defining appropriate *indicators* that are computed from the learners' logs with the aim to reconstitute the actual reading experience. Depending on the analysis goals, these indicators can range from basic statistics like number of visits or visits per web page to more advanced measures like inferring students' attitudes that affect learning (Arroyo et al., 2004) and predicting students' knowledge (Feng et al., 2005).

Traditionally, interaction indicators are calculated using the frequency distribution of the action events recorded on the logs. However, solely relying on an event-based approach to study reading has a major drawback that requesting a page is not necessarily equivalent to reading everything that is presented on it (Hauger et al., 2011). A more robust approach that improves the quality and accuracy of the interpretations derived from the data used, is to take into account the time that elapses between learner events in

addition to the frequency distribution of these events. The use of time between learners' requests improves the quality and the precision of the interpretations that are derived from the data captured on the platform (Kovanovic et al., 2015). However, in the context of server-side monitoring, such an approach is challenging: time cannot be obtained directly from the event stream because web logs contain only action timestamps without explicit markers of their endpoints. A trivial solution is to estimate the duration of each action as the time difference between its occurrence and the occurrence of the next action. The results of such a method could be imprecise since the periods of inactivity can be contained in the estimated durations. Different other methods were proposed to estimate action duration using web log analysis (see (Kovanovic et al., 2015) for an extensive background on time-on-task estimation methods). Another alternative is to identify the active and inactive periods of a user by splitting his action log into "sessions". A session consists of a set of actions assumed to have been performed by the same user in a sustained and continuous activity.

2.3   Session-based indicators in e-learning

Session-based indicators allow to encode the navigation behaviour of users over time (Mobasher, 2007), a valuable aspect that advocates their use to analyse reading efficiently beyond the course and page levels perspectives. When selected appropriately, these indicators provide good insight about learners' behaviour.

The duration and the length of learning sessions (that Marquardt et al. (2004) defined as a set of web sessions) are often used to estimate the success and difficulty of the learning task (Aula et al., 2010) and to reflect readers' level of interest (Sun et al., 2016). Session duration and number of pages are studied in Kazanidis et al. (2012) and used by Thomas (2014) as indicators of learning success. Navigation properties and patterns were studied by Klašnja-Milićević et al. (2011) to give insight into dependencies between page requests. Akçapnar et al. (2010) defined indicators related to rereading within learning sessions to study user disorientation. Nakayama et al. (2000) used page reading order within session indicators to diagnose learning difficulties. However, apart from such general session descriptive metrics common in web-based navigation, session-based indicators related to reading activity are not explicitly addressed. There is even no common definition of the concepts of "session" or "reading session" in e-learning.

2.4   Session identification in e-learning

Methods originated from the field of web usage mining are often used for session identification. This field aims to reveal the knowledge hidden in navigation logs (Munk and Drlik, 2011) and defines two main classes of approaches for reconstructing user sessions: time-oriented and navigation-oriented. The first is based on the limitation of total session time or page-stay time. In the first case, the total duration of a session is limited by a predefined timeout delimiter (threshold), generally 30 minutes (Cooley et al., 1997), and if the duration of accumulated page view-times exceeds that cut-off, the session is classified as having ended. A threshold can also be defined for any page (generally 10 minutes): a session is terminated on a given page if the difference between its access time and the next accessed page is greater than the threshold. A new session is assumed to start with this next accessed page. The navigation-oriented approach uses web topology as a graph and assumes that nodes are web pages and hyperlinks are

directed edges connecting these nodes (Cooley et al., 1999). If a web page is not connected with the previously visited page in a session, then it is considered as contained within a different session.

Without providing precise threshold values, Marquardt et al. (2004) recommended the use of time-based thresholds methods in e-learning. As no generalised model exists for estimating a threshold in a given situation (Huynh and Miller, 2009), authors rely on their data corpus and context characteristics to define appropriate values: 30 min (Del Valle and Duffy, 2009), 60 min (Wise et al., 2013) or even 7 hours (Perera et al., 2009). In e-learning, some authors propose specific approaches like estimating the needed time for reading using an average reading time (in words per minute) (Brown and Green, 2009) or learners self-reports on the time spent (Romero and Barbera, 2011).

3   Log-based analysis of reading activity

3.1   A new approach for identifying sessions of reading activity

In order to study reading activity using the logs collected on learning platforms, we define the concept of "*reading session*" to denote the active period during which this activity takes place. A *reading session* can be defined as *a delimited and sustained set of pages visited and supposedly read by the same reader within the duration of one particular visit to a particular website*. It thus refers to the set of consecutive reading actions from a learner that can be considered continuous (apart from small interruptions, e.g. for reading email). This means that a learner who actually spends one-hour time on a course will carry out a one-hour reading session. Similarly, this concept was used in former studies to characterise reading, for instance on Wikipedia (Lehmann et al., 2014).

In the context of online courses, the navigation-based methods for identifying sessions is not suitable since navigational links may exist between all pages that constitute a course. The time-based approaches are thus more appropriate; yet, the transposition of e-learning characteristics into the web usage mining application is not a trivial task (Zaïane and Luo, 2001). Nevertheless, we identified three major disadvantages for using a single threshold value, whether for the page stay or for the session duration:

1   *E-learning activities are diverse*: Web-based learning is the orchestration of diverse activities (reading, searching, commenting, doing assessments, etc.). Depending on the underlying difficulty, some activities are easier to perform and hence take much less time than others. The existing solutions however do not make distinction of the different learning tasks.

2   *Activity context may change*, a context being related to the learning activity (e.g., time needed to make an assessment depends on the questions difficulties, navigating within a learning portal may be more time demanding than a news one). Hence, each website (course) being unique should have its own time thresholds (Munk and Drlik, 2011).

3   *Pages of the same course website are different*: as educational websites may have complex structures and contents, different difficulty levels for reading and understanding are induced to their pages (introductory parts may be easier to read

and understand than more complex ones). As a matter of fact, each page (containing for instance one chapter or one part) of the same course is different (with regard with its inner-complexity) and thus requires a dedicated reading time.

We propose a new session identification approach where sessions are delimited more efficiently by:

(1) Considering solely reading activity;

(2) Computing page per page stay time and thresholds; and

(3) Using actual time spent by learners reading each page.

This approach defines a dynamic process for estimating action duration and delimiting sessions by using course pages reading-time thresholds. Each page is associated an individual threshold value, computed as the maximum time that was spent by learners to read that page. These values are dynamic since they can evolve when new reading logs are captured and processed.

Technically, the approach is composed of five consecutive steps, with the first two ones (*user identification* and *actions duration estimation*) as pre-processing steps. In the remainder of this paper, we will use *action* as a shortcut for *reading action*. The different phases that compose the session identification algorithm is represented in Figure 1.

Figure 1   Steps for computing reading sessions from learners' log data
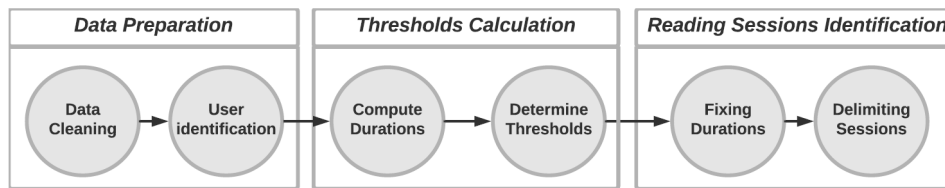
3.1.1 Data preparation

The data gathered by the logging features of the learning environments often contain many irrelevant information. Data preparation refers to a set of pre-processing tasks that are performed on learners' logs to transform them into a suitable format for an easy and effective analysis. Common cleaning tasks are first performed on the raw data in order to detect and remove possible errors and inconsistencies to improve the data quality. Each record on the resulted data is associated with the corresponding learner.

Most modern web servers use the session concept to maintain persistent communication with their clients. For instance, they can associate a unique identifier to each client process accessing the server during all the client visit. In our approach, we use this data as a means to identify unique users. If the user identification is available, we reconstruct for each user his set of requests. If we lack this information or if we suppose that the identification is not required, we assume that each web session is connected to a dedicated anonymous user, each anonymous user being different from the others.

3.1.2   Thresholds calculation

Reading logs refer to the ordered set of time stamped requests representing user interactions with the system. Because the explicit end time of users' actions (an action being between two consecutive requests of the same user) is not captured by server-based logging systems, actions duration is not directly available. Therefore, we use the time order in requests from a given user to assign user's actions end times and durations. For each sequence of actions of a given user, the begin time of each of his request is considered as the end time of his previous action to compute the assumed duration of the action.

Server-based monitoring can result in the recording of very long events, up to several days for parts that can be read in few minutes. This is because a learner may access a course part then change his activity momentously, for a long time or definitively. Moreover, some events may be very short and hence not correspond to actual reading actions. To minimise the impact of these actions on the threshold calculation, we solely use "normal actions", excluding duration-excessive and duration-insignificant actions. We apply *Peirce's criterion*, a method that eliminates the presence of several suspicious data values (outliers) (Ross, 2003). The maximum value of the subset of the data obtained after removing the outliers is hence taken as the element reading threshold. This threshold is used for delimiting reading sessions.

3.1.3   Reading sessions identification

As already explained, we compute the duration of each course element as the time interval between the begin time of that action and the begin time of the action that follows. Unknown durations occur for the last action since no other request can be used to define its end time. In order to not affect the data corpus, and rather than skipping these actions, we assign them with the threshold values of the elements that have been visited with these actions.

We use the reading thresholds of the course elements to split the logs of each learner into sessions. A reading session is assumed finished when the time spent for reading an element is greater than the time threshold of that element. The element that follows that last element is therefore assumed to belong to a different session.

3.2   Reading session-based indicators for analysing reading

We have defined several indicators that are computed using data about learners' reading sessions, originated from widely used metrics in navigation analysis. The complete set of indicators are organised into four classes (see Table 1), and is intended to help characterize reading behaviour from the following perspectives:

1   Learners' interest and their reading pace.

2   Learners' navigation within the course, including how they navigate to and from each course element.

3   Learners' revisitation usages within the course.

4   Learners' reading interruptions, stops and resumes.

Table 1 Definition of the proposed indicators, grouped by classes

Stickiness	
Visits ratio	Number of visits on the element, divided on the count of all visits to the course.
Readers ratio	Number of unique readers of the element, divided on the count of all unique readers of the course.
Reading sessions	Number of learners' reading sessions that contain the element, divided by the count of all the sessions observed on the course.
Reading speed	Size of the element in words, divided by the average time in minutes spent by learners reading the element.
Rereading	
Rereads	Number of revisits divided on the total number of visits.
Within-session reread	Number of revisits within the same sessions divided on the total number of revisits.
Between-session reread	Number of revisits within different divided on the total number of revisits.
Navigation	
Navigation linearity	Number of transitions that come from the element situated just before or that go to the element situated just after, divided by the total number of transitions that includes the element.
Arrival linearity	Number of arrivals from the element situated just before, divided by the number of arrivals to the element.
Futures incoming	Number of departures to distant future elements, divided by the number of departures from the element.
Past incoming	Number of arrivals from distant previous elements, divided by the number of arrivals to the element.
Departure linearity	Number of departures to the element that directly follows, divided by the number of departures from the element.
Future outgoing	Number of departures to distant future elements, divided by the number of departures from the element.
Past outgoing	Number of departures to past elements, divided by the number of departures from the element.
Stops and resumes	
Reading halts ratio	Percentage of reading sessions that are terminated on the element.
Final stops	Percentage of readers who stop reading the course on the element.
Resume linearity	Percentage of reading halts that occur on the element with resumes on that element or the one that follows (normal cases of resume).
Resume to past	Percentage of reading halts that occur on the element with resumes on previous elements.
Resume to future	Percentage of reading halts that occur on the element with resumes on elements far ahead from the current and the direct following ones.

3.2.1 1st class: stickiness and interest indicators

The distribution of the reading sessions over several dimensions can highlight many facts about learners' readings and provide basic hints for readings characterisation within

reading sessions. The stickiness class reflects the popularity of course elements and their ability to attract and hold learners interest. We propose within this class to compute for each course element its ratios of course visits that targeted the element, the unique learners who visited the element, and the ratio of reading sessions that contain the element. We also estimate for each element its average reading speed, expressed in words per minute.

3.2.2   2nd class: reading paths and transitions

Despite the hypertext construction of a course, its elements are often organised in a linear logic to depict the semantic organisation of ideas within the course plan. A linear navigation corresponds to a reading that strictly follows the course plan. The navigation linearity indicates whether the order of reading corresponds to the same order defined by the course plan. This order, tightly related to comprehension (Hahnel et al., 2016), characterises the deviation of the reading paths from the author's expected one. An important deviation often indicates possible readers' disorientation and cognitive overload. The reading path of a user is the sequence of elements that have been read within his reading sessions. Navigation properties were found to be correlated with learning task success (McEneaney, 2001). For instance, the user path and its deviation from the optimal one are used to predict a possible user disorientation (Gwizdka and Spence, 2007). A user transition between two course elements allows reporting a relation between them. This class describes the conformance and deviation of the expected reading path represented by the document plan. The indicators navigation linearity, arrival linearity and departure linearity aim to characterise how learners navigate to and from each course element. To deepen the analysis, we defined other indicators that allow to examine how the linearity is being broken at (1) arrival: from elements situated after the studied element (future incoming) or before the element that proceeds (past incoming); and departure: to course elements situated after (future outgoing) or far before (past outgoing).

3.2.3   3rd class: rereading

Rereading is a common browsing behaviour and a strategy used spontaneously by struggling readers (Akçapnar et al., 2010; Wise et al., 2012). We define two types of rereading: *within-session rereads* occur on the same reading session, and *between-session rereads* are performed on different sessions.

   In general, revisiting previously seen content might reflect intellectual processing difficulties (Hyönä et al., 2003). When a given chapter is being intensively reread, this often means that learners need from the author to better clarify the discourse to make it more accessible. This is even true when the rereads are within the same sessions. The rereads situated in different sessions can be seen as the expression of content complexity and memorisation difficulties. That means that learners need reminders of the earlier visited elements (e.g., to understand new concepts presented later).

3.2.4   4th class: reading stop and resume

Analysing reading session interruptions allows to understand how learners interrupt and resume reading. Some interruptions are final (reading *final stops*), meaning that the user

no longer returns to complete the reading of the course. While some cases are trivial (e.g., last chapters of the course), other cases can indicate that learners lost motivation and interest on the course. No final stops (*reading halts*) are followed by resumes on normally either the same element or on the next one (*linear resume*). Important abnormal resumes indicate that learners need to navigate elsewhere to understand the presented information.

4 Evaluation methodology

4.1 Evaluation context and objectives

Our proposals are implemented and evaluated using data from *OpenClassrooms* (http://www.openclassrooms.com/), a major French e-learning platform that provides courses in information technology, entrepreneurship, and digital skills. With more than 1000 courses and 1 million members, OpenClassrooms totalises about 2.5 million unique visitors every month. Course authors are generally domain-experts; some of them are academic teachers and instructors. Courses are organised as the nesting of *course elements* (corresponding to parts and chapters), each element being contained in a dedicated webpage. Logfiles contain information about website visitors' activity and are automatically created by the web server. Common cleaning and pre-processing steps are performed to obtain for each record a *timestamp* (datetime of the request) along with the *request identifier*, the *user* (empty if anonymous), the *server-side session*, the *course* and the *course element*.

A record within the data corpus has the following structure:

```
<request_id, user_id, course_id, element_id,
server_session_id, timestamp>
```

The corpus used is composed of twelve courses among the most popular on the platform. Data used are learners' logs for the period starting from *31 October 2014* to *07 July 2016*. Table 2 provides statistical data about these courses. As the standard deviation (SD) values indicate, there's evidence of distribution inequality of these variables within the different courses.

We evaluated our proposals using different methods with different objectives. The items under scrutiny were:

- The capabilities of the proposed session identification algorithm to detect reading sessions that are compliant with learners' actual ones (*RQ*1).

- The perceived relevance of the proposed indicators for course revision, from the authors' perspective (*RQ*2).

- The effectiveness of using the proposed indicators for detecting reading issues and revision needs, from the learners' perspective (*RQ*3).

In the next paragraphs, we describe the methodology used for evaluating each of these three items, along with the results and main findings.

Table 2   Basic statistics about the selected courses

Course	# chapters	# logs	# learners	# reading sessions
Bootstrap	7	229,362	13,045	94,654
Web	18	240,978	11,793	53,695
Twitter	9	17,576	1560	5223
Adruino	14	66,911	4864	26,797
JavaScript	13	289,153	12,829	101,614
Ionic	19	27,283	2020	8663
Ruby	18	4895	706	2794
Project management	14	49,255	3156	14,607
TCP	17	111,026	7239	43,392
Symfony	27	402,039	9357	236,635
Startups	21	11,772	1223	3574
Github	19	109,092	5826	29,452
Median	17.5	88001.5	5345	28124.5
Mean	16.33	129945.2	6134.833	51758.33
SD	5.38	129909.1	4653.657	67299.6

4.2   Reading session calculation (RQ1)

4.2.1 Methodology

In the absence of a precise knowledge of the actual sessions that compose the navigation traces of the readers, it is in practice impossible to verify with confidence the conformance of the reconstructed sessions. Many researchers have previously studied the properties of such sessions and found that their size in terms of total number of the visited pages follows a *power law* distribution (Berendt et al., 2001; Vázquez et al., 2006). As no effective measurement can assess the compliance of the reconstructed sessions with the actual ones, we propose to study the reconstruction quality of our method using this empirical observed power law distribution. More precisely, this law states that most visits to a website are concentrated on a small number of pages, with the rest of the pages receiving a less important number of visits. Such an approach was also used for the same purpose of analysing the quality of session reconstruction by many authors (e.g., Arce et al., 2014; Dell et al., 2008; Román et al., 2014). We also study the compliance of the computed action durations with course element complexity with the assumption that complex course elements take more time to read and understand. We deal with complexity by computing the element size, which is one of the numerous factors that allow to characterise a website complexity (Butkiewicz et al., 2011). We also compare the results of our approach with popular web usage mining methods, in terms of quality of reconstruction, and compliance with course element complexity.
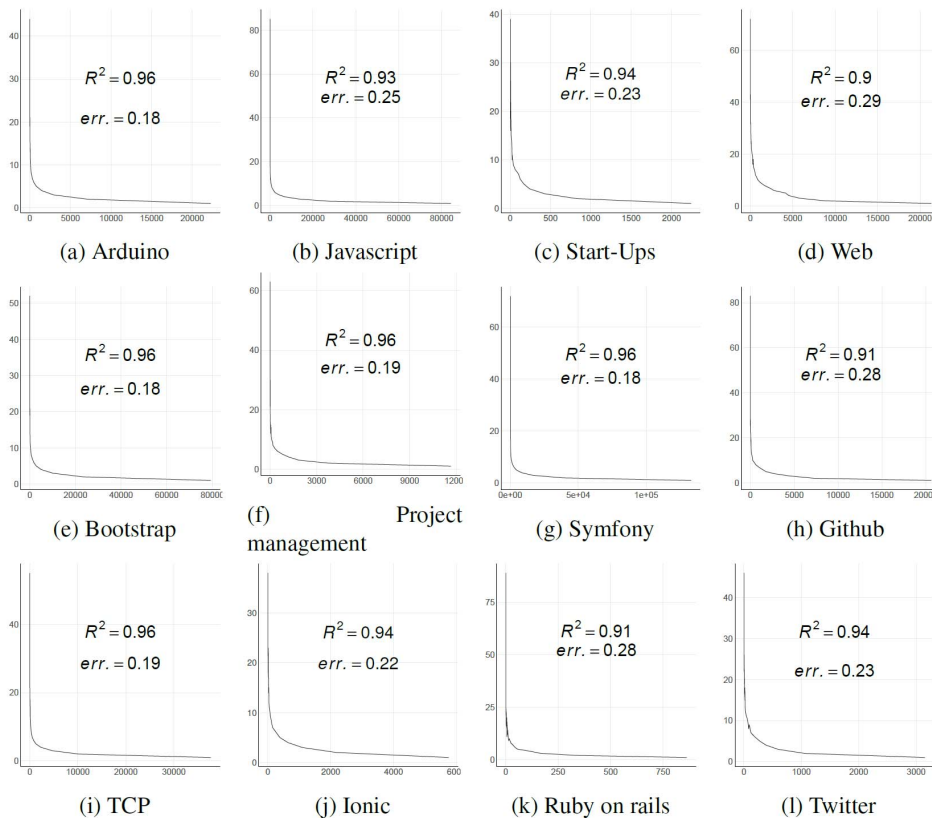
4.2.2 Results

4.2.3 Quality of the reconstruction

Evaluating the quality of reconstruction using the power law is generally performed using a linear regression on the logarithm of the number of the distinct read elements and the logarithm of the total number of reading sessions. The quality measure is given by the regression correlation coefficient R^2 and the standard error *err*. The closer R^2 coefficient is to one and the *err* near to zero, the better the session identification result.

We applied the proposed session detection approach on the twelve courses; we then used the power law to estimate the quality of the reconstruction. According to the results, represented on Figure 2, the approach exposes good capabilities for our context of study since it gives good fit results with a relatively good accuracy for all the courses. Indeed, the values of the different regression correlation coefficients R^2 are all above or equal to 0.90 (mean value of 0.94). The different values of the standard error *err* are also acceptable (mean value of 0.22).

Figure 2   Session size found by the power law distribution on the 12 courses

(a) Arduino   (b) Javascript   (c) Start-Ups   (d) Web

(e) Bootstrap   (f) Project management   (g) Symfony   (h) Github

(i) TCP   (j) Ionic   (k) Ruby on rails   (l) Twitter

In order to reinforce our results, we recomputed the learners' sessions on the 12 courses using two other methods. We first used a fixed value of threshold on the page stay time; we chose 10 minutes, given that it is the most used value for this class of approaches. We

also used a fixed value of threshold on total session time; we used 30 minutes since this is the widely used value for this class. We then estimated the quality of the reconstruction for each case, in order to compare the results of these approaches with ours.

The results of the comparative study are given on Table 3, we emphasise in bold the best R^2 coefficient and *Err* error values. The method we proposed gave best fit results for the majority of the courses with an acceptable accuracy given by the error values. More precisely, it gave (1) best fit result for 75% of the courses (9 courses out of the 12); and (2) best values for the fitness and accuracy couple in 58% of cases (7 courses out of the 12). One side result of this study is that, for our study context, the use of thresholds on page stay time (fixed or dynamic) seems more effective for delimiting reading sessions than the use of a fixed total session duration.

Table 3   Constructed sessions using three methods: our proposal, fixed page threshold (10-min) and fixed session threshold (30-min).

	Reading Session		10-min Page Threshold		30-min Session Threshold	
	R^2	Err	R^2	Err	R^2	Err
Bootstrap	**0.96**	**0.18**	0.95	0.22	**0.96**	0.21
Web	**0.90**	**0.29**	0.88	0.31	0.86	0.31
Twitter	**0.94**	**0.23**	**0.94**	0.24	0.91	0.25
Adruino	**0.96**	**0.18**	0.94	0.20	0.93	0.23
JavaScript	**0.93**	**0.25**	0.92	0.26	0.90	0.28
Ionic	**0.94**	0.22	0.93	**0.21**	0.92	0.24
Ruby	0.91	0.28	0.91	0.29	**0.93**	**0.26**
Project management	**0.96**	**0.18**	0.92	0.24	0.92	0.23
TCP	**0.96**	**0.19**	0.95	0.20	0.95	0.20
Symfony	**0.96**	**0.29**	0.88	0.31	0.86	0.35
Startups	**0.94**	0.23	0.93	**0.22**	0.93	0.23
Github	0.91	0.28	**0.94**	**0.23**	0.93	0.25

4.2.4 Compliance with elements size and complexity

We estimated the size of each element of the courses by counting its significant words and in-line images (with each image considered as a short paragraph of 30 words). Pearson correlation coefficient between element size (computed as the words and figure count) and time threshold for that element is $r = 0.82$ ($p < 0.001$). This positive and significant correlation means that our method is actually generic and robust enough to take into account element size without needing to calculate it for each element. We can make the hypothesis that it is also the case for element complexity, even if element size does not directly indicate complexity level of the content. This also confirms the need to take into account characteristics of the elements for more accurate threshold values.

4.2.5 Comparison with fixed threshold methods

Following a per page reading time threshold approach for delimiting reading sessions based on learners reflects the actual usages and differentiates elements based on their content. Using a sole fix value would give imprecise results since we assume for each

element the same maximum reading time. In fact, some elements may be read faster while others may need more than the fixed value time for reading. As can be seen for the TCP course on Figure 3(a), using a sole threshold for page stay results in sessions that are different from the actual ones. Figure 4 shows for each course the distribution of reading time for each of its chapters. Whatever the selected threshold, there will always be elements which can be read in less than this time and others that may take more time. Our method does not define constraints on the session size, which would cut many continuous long sessions and merge other short ones, as illustrated on Figure 3(b) for the TCP course.

4.3   Relevance of the proposed indicators for analysing course reading(RQ2)

4.3.1   Methodology

We conducted an exploratory user study with course authors in order to evaluate the relevance of the four classes of indicators for analysing and planning revisions. In this perspective, we designed an online questionnaire that first describes the project rationale, then presents the set of indicators and illustrates their possible use cases. About two hundred authors with one or several courses available on the Openclassrooms platform were contacted by mail and asked to fill out the questionnaire. We received a total of 105 responses.

Using the questionnaire, the authors were asked to read the explanation of each indicator and to rate using *Likert scales* (1 = not useful, 5 = very useful) the usefulness and relevance of each indicator for (1) gaining more insight about the course reading behaviour; and (2) for detecting possible issues that may hamper learners from understanding the presented content. The authors were free to discuss the indicator classes and to justify their own ratings within a comment section.

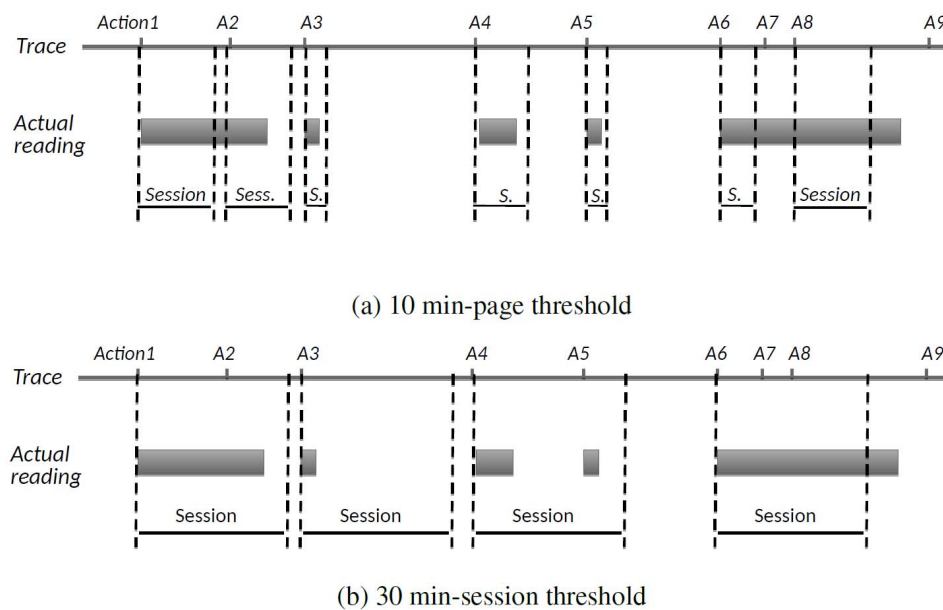**Figure 3**   Constructing sessions using time-based heuristics (TCP course)
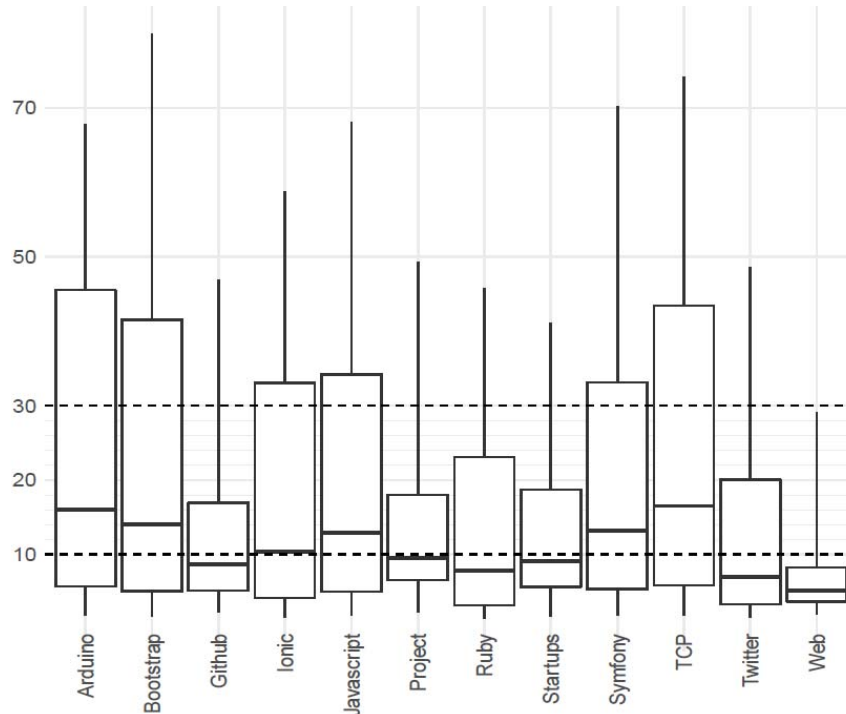

(a) 10 min-page threshold

(b) 30 min-session threshold

Figure 4   Course element reading duration statistics (boxplots)


4.3.2 Results

There are two schools of thoughts on analysing Likert-scale data: ordinal versus interval (Carifio and Perla, 2007). A significant amount of empirical evidence supporting that Likert scales can be used as interval data exists (Carifio and Perla, 2007). Moreover, according to (Cooper et al., 2006), the Likert-scale items that measure the same variables can be summed up to create new composite metric scale. Accordingly, we aggregated the individual indicator ratings into their corresponding classes. The Cronbach α values of the results range from 0.78 to 0.91, which validates the internal consistency and reliability of the responses.

We analysed the results of the questionnaire using descriptive statistics. The descriptive statistics of the aggregated results are presented on the first three columns of Table 4 and the boxplots represented on Figure 5. The boxplots are relatively short, suggesting that the participants had a high level agreement with each other. The results show that participants mostly agreed that the classes of indicators are relevant for analysing course reading. All the proposed classes of indicators were highly rated; indeed, the median and mean points were all above the neutral point of 3 (corresponding to the mention "neither agree nor disagree"). In other words, globally the four classes of indicators were deemed useful. The navigation indicators were deemed less relevant. The ratings of the individual indicators showed that the more contrasted results are related to ratio of visits, rereading ratio, and stops without resumes. Indicators related to the *average reading speed* and *session ratios* were the lowest rated ones.

Table 4   Statistics about authors' ratings

	Descriptive statistics			Student t-test[*]			Inter-correlations (Spearman)		
	Mean	*SD*	*Median*	*t*	*df*	*p*	*Navigation*	*Rereading*	*Stops*
Stickiness	3.66	0.53	3.71	0.41	105	0.67	0.01	0.16	0.13
Navigation	3.44	0.54	3.33	0.47	105	0.64		−0.01	−0.04
Rereading	3.67	0.59	3.67	−0.04	105	0.96			0.01
Stops	3.55	0.60	3.67	−0.89	105	0.33			

Note: [*]group 1: female participants ($n = 42$); group 2: male participants ($n = 63$)

Figure 5   Authors' rating of the indicators, aggregated by classes
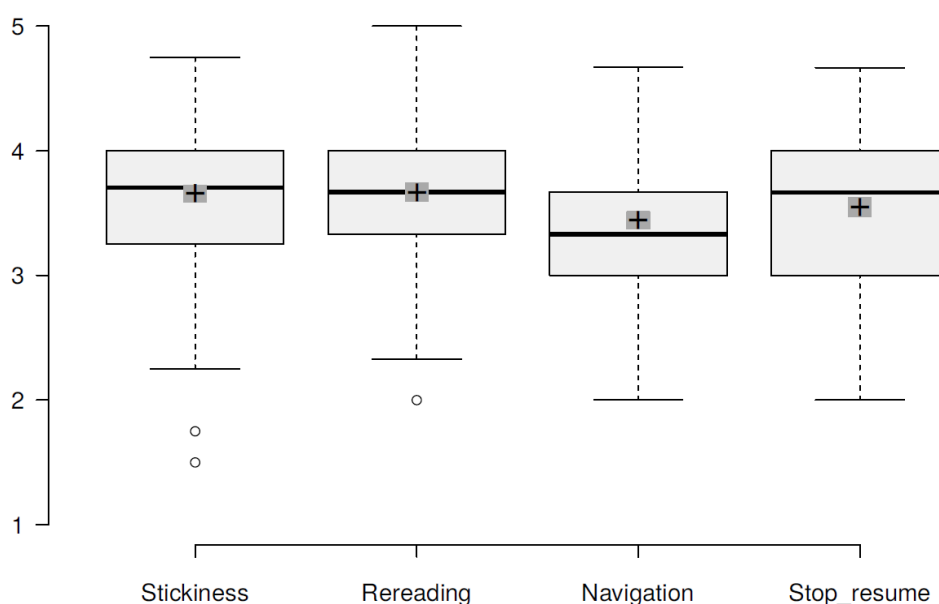
In order to investigate for possible differences between the genders, we ran an independent *t-test* for each class of indicators (the significance level for the mean variation was set at $p < 0.05$). The results, shown on Table 4 (columns 5, 6 and 7), revealed no significant difference for all the classes of indicators. Associations between ratings of the different classes of indicators were examined using *Pearson's correlation coefficients*. As reported on the last columns of Table 4, no significant correlation was found.

According to authors' comments, these indicators were found judicious to give a good idea of the way learners read courses. The following sample quotes illustrate some authors' opinion:

• "What is interesting is this opportunity given to the authors of online courses and their learning to communicate, in a direct and indirect way."

• "Why not to include direct exchanges between authors and readers through comments and forums?"

- "These are important metrics about course consumption; they could help me understand how to rethink my course material."
- "While they seem interesting, I think you would have to select the more important indicators to present to authors. The other ones can serve for deeper analysis."
- "Be careful not to abuse the personal data of users. The reader should actually be informed that his reading is logged and analysed."

While more than 73 authors estimate the set of indicators comprehensive enough to analyse reading, eight authors think that there are too many and without a judicious presentation to authors, this would be counterproductive. An author estimates that the approach seems complicated to implement technically and therefore may generate some unreliable results. Similarly, another author believes that we need a good level of abstraction so that authors will not be required to consult many tables and endless statistics. The authors also discussed privacy concerns; they suggested asking learners before logging them. Several authors proposed to consider the supplementation of computed indicators with explicit learners' feedback (courses and elements ratings, comments and annotations, etc.) that would help them to fully understand learners' needs.

4.4   Effectiveness of the detected issues (RQ3)

4.4.1 Methodology

By examining the values of the indicators for the different course elements, it is possible to deduce that some elements may present some aspects and content that hamper learners' understanding. The purpose of this part of the evaluation is to examine whether the detected reading issues correspond to real comprehension difficulties encountered by learners.

Participants in this part of the study were Algerian master's students in Computer Sciences and Information Systems who enrolled in an advanced training in Information Systems at the Algerian *Research Centre on Scientific and Technical Information* (CERIST). A total of 26 master's students (10 female and 16 male, from 23 to 26 years old) took part to the study.

Among the twelve courses we have already analysed, the students selected four courses that they indicated have already followed on the platform during the first and/or the second semester of the year 2017 (*TCP*, *JavaScript*, *Bootstrap* and *Symfony*). The reading logs of these courses, provided by OpenClassrooms, were used to compute the values of the different indicators. We examined the statistical distribution of the values of these indicators for each course, to assess possible reading issues. The rules we followed for marking a given value of a specific indicator as reflecting a potential reading issue are provided on Table 5. The numerical results of the issue detection on the four courses are shown on Table 6.

Table 5   Issue detection using indicator value

Class	Issue triggering	Issue description
Stickiness	Low values of: visits, readers, reading session; high reading speed	(SI_1) Low popularity due to low attractiveness of the chapter and/ or its low readability
	Low reading speed	(SI_2) Low stickiness due to the complexity of the content
Navigation	Low linearity of reading (arrivals and/or departures); high ratios of navigation to distant chapters	(NI_1) Disorientation due to bad structuring
		(NI_2) Non-linear reading due to low memorability
		(NI_3) Non-linear reading due to low content complexity
Rereading	High values of rereads and/or within-session rereads	(RRI_1) Many consecutive rereading due to content complexity
	High values of between-session rereads	(RRI_2) Many distant rereading, due to low memorability
Stop and resume	High values of final reading stops	(SRI_1) Permanently stop reading the course because of loss of interest, poor readability and/or high complexity
	High values of reading halts (non-final stops)	(SRI_2) Reading halts due to content complexity
	High values of non-linear resume; high values of resume on distant chapters	(SRI_3) Resuming on previous on future distant chapters due to content complexity, low memorability and/or bad structuring

Table 6   Main reading issues detected on four courses

	Stickiness		Navigation			Rereading		Stop	and resume		ALL
Course	SI_1	SI_2	NI_1	NI_2	NI_3	RRI_1	RRI_2	SRI_1	SRI_2	SRI_3	
TCP	3	1	2	3	0	1	2	1	0	1	14
Javascript	1	1	1	1	0	1	0	1	0	2	8
Symfony	3	1	2	1	1	2	1	2	1	1	15
Bootstrap	1	1	1	2	1	2	2	0	1	2	13

After gathering basic demographic characteristics of the students, we presented them a paper-based questionnaire that listed for each class of indicators the issues we assessed, supplemented with summary explanations. We asked the students to carefully examine the marked issues and to rate their effectiveness using five-points *Likert scales* (1 = absolutely disagree, 5 = absolutely agree).

4.4.2 Results

The descriptive statistics of the results are shown on Table 7. The distributional characteristics of the ratings are also represented on Figure 6. All the issues had good rating values, with a median that is superior to the neutral point of 3 except for SRI_3 which median is equal to the neutral point. This suggests that the participating students

acknowledged that most of the detected issues correspond to real problems within the course that may hamper easily reading it and understanding its ideas. Issues related to course element popularity $(SI1)$ and content complexity (RRI_1, SRI_1 and NI_1) were the most highly rated.

Table 7   Descriptive statistics about learners' rating of the effectiveness of the issues (1 = very low, 5 = very high), and *t-test* results based on gender difference

	Descriptive statistics			Student t-test* results		
	Mean	SD	Median	t	df	p
Stickiness issues						
SI_1	4.04	0.96	4.00	−0.57	24	0.57
SI_2	3.15	1.26	3.00	−0.81	24	0.43
Navigation issues						
NI_1	3.58	0.86	4.00	−1.32	24	0.20
NI_2	3.50	1.03	4.00	−1.18	24	0.25
NI_3	3.42	1.03	4.00	−1.28	24	0.21
Rereading issues						
RRI_1	3.96	1.04	4.00	−1.88	24	0.07
RRI_2	3.39	1.02	3.50	−0.33	24	0.75
Stop and resume issues						
SRI_1	3.73	1.00	4.00	−1.35	24	0.19
SRI_2	3.46	0.95	4.00	−1.54	24	0.13
SRI_3	3.31	1.12	3.00	−0.74	24	0.47

Note: *group 1: female participants (n = 10); group 2: male participants (n = 16)

Figure 6   Learners' rating of the effectiveness of the issues (1 = very low, 5 = very high)
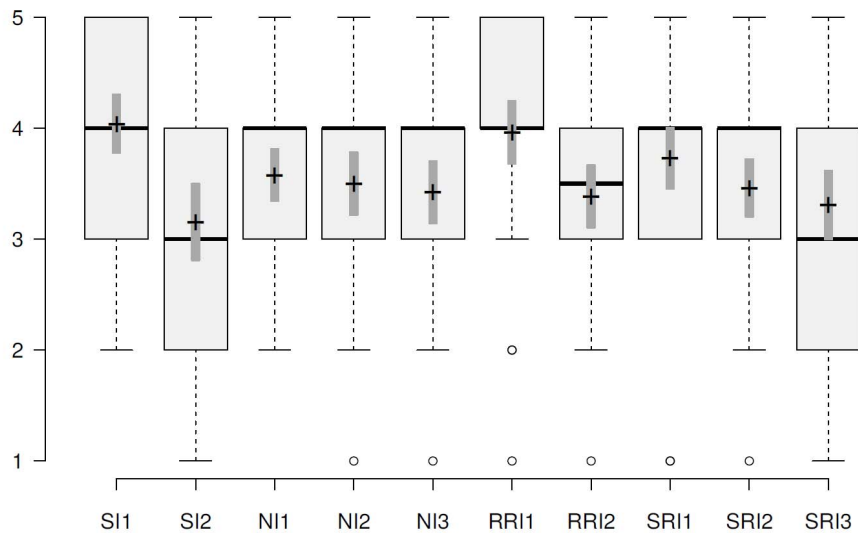
To examine the difference in ratings between male and female participants, we conducted an independent-samples *t-test* analysis (with the significance level for the mean variation set at $p < 0.05$). The results (the last three columns of Table 7) show no statistically significant difference between the two groups (female versus male). There is a clear difference in the rating distribution for the different indicators in terms of skewness.

To further investigate the results, we used Pearson correlations between the issue ratings to determine if any association existed between them. The results shown on Table 8 reveal many significant relations. There are moderate negative correlations between score of stickiness issues and issues related to reading halts and non-linear resumes (SI_1 and SRI_3 with $r = -0.46, p < .05$; SI_2 and SRI_2, with $r = -0.50, p < .01$). This suggests that the failure to attract learners' interest is also reflected by the tendency of learners to stop or to momentously interrupt reading. There are also moderate to strong positive correlations between rereading and reading stops and non-linear resume issues RRI_1 and SRI_2, with ($r = 0.64, p < .001$; RRI_1 and SRI_2, with $0.55, p < .01$; RRI_1 and SRI_3, with $0.64, p < .05$; RRI_2 and SRI_1, with $0.57, p < .001$). The majority of these issues reflect the difficulty for learners to grasp meaning. Positive moderate inter-correlations within class issues exist: issues related to stops and resume (SRI_1 and SRI_2, with $0.64, p < .001$; SRI_1 and SRI_3, with $0.54, p < .01$), navigation issues (NI_1 and NI_2, with $0.52, p < .01$; NI_2 and NI_3, with $r = 0.47, p < .05$), and rereading issues (RRI_1 and RRI_2, with $0.69, p < .001$). Finally, we found a weak correlation between two issues (NI_1 and SRI_3, with $r = 0.39, p < .05$), both of them reflect possible disorientation due to the course structuring.

Table 8   Inter-correlations (Spearman) among learners' ratings of the different types of issues

	SI_2	NI_1	NI_2	NI_3	RRI_1	RRI_2	SRI_1	SRI_2	SRI_3
SI_1	–0.30	–0.13	–0.02	0.23	0.01	–0.18	–0.19	0.16	–0.46*
SI_2		0.04	0.28	0.08	–0.18	–0.17	–0.09	–0.50**	0.08
NI_1			0.52**	0.21	0.30	0.24	0.37	0.15	0.39*
NI_2				0.48*	0.35	0.38	0.36	0.20	0.31
NI_3					0.20	0.22	–0.01	0.24	0.02
RRI_1						0.70***	0.64***	0.55**	0.42*
RRI_2							0.57**	0.51	0.24
SRI_1								0.64***	0.54**
SRI_2									0.35

Note: *p<.05, **p<.01, ***p<.001

5   Discussions and conclusions

Our first proposal in this paper is a new method for delimiting learners' reading sessions by computing page per page thresholds. The proposed method is grounded on data that represent learners' interactions with course elements and that take into account each page

characteristics. The resulting thresholds are dynamic since their values are recalculated each time new reading actions are logged. This allows their automatic updating to (1) adjust their values to incoming reading data, and to (2) take into account any evolution of the courses like pages restructuring and content update.

According to the evaluation results, the proposed dynamic method allows to better simulate learners reading, and to fit the expected statistical behaviour of real sessions. The results also consolidated our statement that the use of fixed-value methods (for session duration or for page stay time) may not be appropriate for educational websites. Indeed, unique threshold values are not suitable for considering neither the specificity of the courses nor the content of their different pages. We plan to further verify the method capabilities by defining more accurate metrics to characterise elements complexity and to compare the deduced session's compliance with the real ones.

Modelling reading activity using sessions allowed us to define indicators that describe the underlying process from behavioural perspectives. We have proposed several reading indicators that are constructed and calculated using data about learners' reading sessions. Their aim was to better represent and explain how learners consume and assimilate the content offered on the learning platform. For the authors that participated in the evaluation study, the exchanges between course authors and learners are an essential aspect within the teaching and learning practices to build interesting and productive pedagogical contents. They globally acknowledged the usefulness of the proposed indicators and confirmed their relevance to guide them in improving their courses.

We used reading indicators not only to describe the reading behaviour of learners, but also as clues to alert about the comprehension problems that these learners might encounter. To achieve this objective, an analysis of the different values of a given indicator could indicate which chapters to examine and which aspects of these chapters should be improved. The study we conducted with learners showed that this approach and its subsequent implementations offer an effective way of reflecting possible reading problems related to the structure and content of the course. Such information can make the author aware of the difficulties of comprehension encountered by their learners, and encourage them to think about how to adjust their courses to make them easier to read and understand.

Despite the encouraging results of the evaluation, we must stress out that these studies were exploratory given the limited number of participants and the controlled settings of the study. Moreover, the learning took place in informal settings, and the courses were delivered through a self-directed e-learning platform. To be able to confirm and generalise our conclusions, we think that broader studies that involve more participants are mandatory. Most of the gathered comments and suggestions from the participating course authors and learners correspond to aspects of our future work that we will address within our main project towards usage-based document reengineering: to further precise the reading indicators, and supplement these with readers' annotations and build a simple and intuitive dashboard to analyse course authors and to assist them evaluating and improving their courses. By deploying such a tool, it would be possible to better study and evaluate the impact of our methodology on learning outcomes.

References

Akçapnar, G., Co s gun, E. and Altun, A. (2010) 'Prediction of perceived disorientation in online learning environment with random forest regression', in Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C. and Stamper, J. (eds): *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, Netherlands, pp.259–264.

Arce, T., Román, P.E., Velásquez, J. and Parada, V. (2014) 'Identifying web sessions with simulated annealing', *Expert Systems with Applications*, Vol. 41, No. 4, pp.1593–1600.

Arroyo, I., Murray, T., Woolf, B.P. and Beal, C. (2004) 'Inferring unobservable learning variables from students' help seeking behavior', in Lester, J.C., Vicari, R.M. and ParaguaÃ§u, F. (eds): *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Alagoas, Brazil, pp.782–784.

Aula, A., Khan, R.M. and Guan, Z. (2010) 'How does search behavior change as search becomes more difficult?', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, Atlanta, GA, USA, pp.35–44.

Berendt, B., Mobasher, B., Spiliopoulou, M. and Wiltshire, J. (2001) 'Measuring the accuracy of sessionizers for web usage analysis', *Proceedings of the Workshop on Web Mining at SIAM Data Mining Conference*, Chicago, IL, USA, pp.7–14.

Brown, A.H. and Green, T. (2009) 'Time students spend reading threaded discussions in online graduate courses requiring asynchronous participation', *The International Review of Research in Open and Distributed Learning*, Vol. 10, No. 6, pp.51–64.

Brown, W.J. et al. (2016) 'Influencing metacognition in a traditional classroom environment through learning analytics', in Li, Y. et al. (eds): *State-of-the-Art and Future Directions of Smart Learning. Lecture Notes in Educational Technology*, pp.1–12.

Butkiewicz, M., Madhyastha, H. V. and Sekar, V. (2011) 'Understanding website complexity: measurements, metrics, and implications', *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, Berlin, Germany, pp.313–328.

Carifio, J. and Perla, R.J. (2007) 'Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes', *Journal of Social Sciences*, Vol. 3, No. 3, pp.106–116.

Cho, K. and MacArthur, C. (2011) 'Learning by reviewing', *Journal of Educational Psychology*, Vol. 103, No. 1, pp.73–84.

Cocea, M. and Weibelzahl, S. (2011) 'Disengagement detection in online learning: validation studies and perspectives', *IEEE Transactions on Learning Technologies*, Vol. 4, No. 2, pp.114–124.

Cooley, R., Mobasher, B. and Srivastava, J. (1997) 'Web mining: information and pattern discovery on the world wide web', *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, USA, pp.558–567.

Cooley, R., Mobasher, B. and Srivastava, J. (1999) 'Data preparation for mining world wide web browsing patterns', *Knowledge and Information Systems*, Vol. 1, No. 1, pp.5–32.

Cooper, D.R., Schindler, P.S. and Sun, J. (2006) *Business Research Methods*, Oxford University Press, USA.

Copeland, L., Gedeon, T. and Caldwell, S. (2015) 'Effects of text difficulty and readers on predicting reading comprehension from eye movements', *Proceedings of the 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Gyor, Hungary, pp.407–412.

D'Mello, S., Dieterle, E. and Duckworth, A. (2017) 'Advanced, analytic, automated (aaa) measurement of engagement during learning', *Educational Psychologist*, Vol. 52, No. 2, pp.104–123.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S. and Nardy, A. (2014) 'Mining texts, learner productions and strategies with ReaderBench', in Peña-Ayala, A. (eds): *Educational Data Mining. Studies in Computational Intelligence*, Vol. 524, pp.345–377.

Del Valle, R. and Duffy, T. (2009) 'Online learning: learner characteristics and their approaches to managing learning', *Instructional Science*, Vol. 37, No. 2, pp.129–149.

Dell, R.F., Roman, P.E. and Velasquez, J.D. (2008) 'Web user session reconstruction using integer programming', *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia, pp.385–388.

Edwards, R.L., Davis, S.K., Hadwin, A.F. and Milford, T.M. (2017) 'Using predictive analytics in a self-regulated learning university course to promote student success', *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)*, Vancouver, Canada, pp.556–557.

Feng, M., Heffernan, N.T. and Koedinger, K.R. (2005) 'Looking for sources of error in predicting student's knowledge', *Proceedings of the AAAI Workshop on Educational Data Mining*, Menlo Park, CA, USA, pp.54–61.

Fidalgo-Blanco, Á., Sein-Echaluce, M.L., Garca-Peñalvo, F.J. and Conde, M.Á. (2015) 'Using learning analytics to improve teamwork assessment', *Computers in Human Behavior*, Vol. 47, pp.149–156.

Gwizdka, J. and Spence, I. (2007) 'Implicit measures of lostness and success in web navigation', *Interacting with Computers*, Vol. 19, No. 3, pp.357–369.

Hahnel, C., Goldhammer, F., Naumann, J. and Kröhne, U. (2016) 'Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text', *Computers in Human Behavior*, Vol. 55, pp.486–500.

Hauger, D., Paramythis, A. and Weibelzahl, S. (2011) 'Using browser interaction data to determine page reading behavior', in Konstan, J.A., Conejo, R., Marzo, J.L. and Oliver, N. (eds): in *User Modeling, Adaption and Personalization. Lecture Notes in Computer Science*, Vol. 6787, pp.147–158.

Huynh, T. and Miller, J. (2009) 'Empirical observations on the session timeout threshold', *Information Processing and Management*, Vol. 45, No.5, pp.513–528.

Hyönä, J., Lorch, R. and Rinck, M. (2003) 'Eye movement measures to study global text processing', *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, Elsevier, Amsterdam, pp.313–334.

Jo, I-H., Kim, D. and Yoon, M. (2014) 'Analyzing the log patterns of adult learners in LMS using learning analytics', *Proceedings of the 4th International Conference on Learning Analytics and Knowledge (LAK'14)* , Indianapolis, IN, USA, pp.183–187.

Kazanidis, I., Valsamidis, S., Kontogiannis, S. and Karakos, A. (2012) 'Measuring and mining LMS data', *Proceedings of the 16th Panhellenic Conference on Informatics (PCI'12)*, Piraeus, Greece, pp.296–301.

Kizilcec, R.F., Pérez-Sanagustn, M. and Maldonado, J.J. (2017) 'Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses', *Computers and Education*, Vol. 104, pp.18–33.

Klašnja-Milićević, A., Vesin, B., Ivanovi c′, M. and Budimac, Z. (2011) 'E-learning personalization based on hybrid recommendation strategy and learning style identification', *Computers and Education*, Vol. 56, No. 3, pp.885–899.

Kovanovic, V., Gaševic, D., Dawson, S., Joksimovic, S., Baker, R.S. and Hatala, M. (2015) 'Penetrating the black box of time-on-task estimation', *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, Poughkepsie, NY, USA, pp.184–193.

Lehmann, J., Müller-Birn, C., Laniado, D., Lalmas, M. and Kaltenbrunner, A. (2014) 'Reader preferences and behavior on Wikipedia', *Proceedings of the 25th ACM conference on Hypertext and Social Media*, Santiago, Chile, pp.88–97.

Liu, R., Patel, R. and Koedinger, K. R. (2016) 'Modeling common misconceptions in learning process data', *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, Edinburgh, Scotland, pp.369–377.

Marquardt, C.G., Becker, K. and Ruiz, D.D. (2004) 'A pre-processing tool for web usage mining in the distance education domain', *Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04)*, Coimbra, Portugal, pp.78–87.

McEneaney, J.E. (2001) 'Graphic and numerical methods to assess navigation in hypertext', *International Journal of Human-Computer Studies*, Vol. 55, No. 5, pp.761–786.

Mobasher, B. (2007) 'Data mining for web personalization', *The Adaptive Web*, pp.90–135.

Munk, M. and Drlik, M. (2011) 'Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining', in Cherifi, H., Zain, J.M. and El-Qawasmeh, E. (eds): *Digital Information and Communication Technology and its Applications, Communications in Computer and Information Science*, Vol. 166, pp.60–74.

Nakayama, T., Kato, H. and Yamane, Y. (2000) 'Discovering the gap between Web site designers' expectations and users' behavior', *Computer Networks*, Vol. 33, No. 1, pp.811–822.

Pattanasri, N., Mukunoki, M. and Minoh, M. (2012) 'Learning to estimate slide comprehension in classrooms with support vector machines', *IEEE Transactions on Learning Technologies*, Vol. 5, No. 1, pp.52–61.

Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaane, O.R. (2009) 'Clustering and sequential pattern mining of online collaborative learning data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 6, pp.759–772.

Román, P.E., Dell, R.F., Velásquez, J.D. and Loyola, P.S. (2014) 'Identifying user sessions from web server logs with integer programming', *Intelligent Data Analysis*, Vol. 18, No. 1, pp.43–61.

Romero, M. and Barberà, E. (2011) 'Quality of e-learners' time and learning performance beyond quantitative time-on-task', *The International Review of Research in Open and Distributed Learning*, Vol.12, No. 5, pp.125–137.

Ross, S. M. (2003) 'Peirce's criterion for the elimination of suspect experimental data', *Journal of Engineering Technology*, Vol. 20, No. 2, pp.38–41.

Sadallah, M., Encelle, B., Maredj, A.E. and Prié, Y. (2015) 'Towards reading session-based indicators in educational reading analytics', in Conole, G., Klobucar, T., Rensing, C., Konert, J. and Lavoué, E. (eds): *Proceedings of the 10th European Conference on Technology Enhanced Learning (EC-TEL'15)*, Toledo, Spain, pp.297–310.

Siemens, G. (2012) 'Learning analytics: envisioning a research discipline and a domain of practice', *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, Vancouver, Canada, pp.4–8.

Siemens, G. (2013) 'Learning analytics: the emergence of a discipline', *American Behavioral Scientist*, Vol. 57, No. 10, pp.1380–1400.

Snodgrass Rangel, V., Bell, E.R., Monroy, C. and Whitaker, J.R. (2015) 'Toward a new approach to the evaluation of a digital curriculum using learning analytics', *Journal of Research on Technology in Education*, Vol. 47, No. 2, pp.89–104.

Sun, J.C., Lin, C.T. and Chou, C. (2016) 'Applying learning analytics to explore the influence of online learners motivation on their online learning behavioral patterns', *Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics*, Kumamoto, Japan, pp. 377–380.

Tempelaar, D.T., Rienties, B. and Nguyen, Q. (2017) 'Towards actionable learning analytics using dispositions', *IEEE Transactions on Learning Technologies*, Vol. 10, No. 1, pp.6–16.

Thomas, P. (2014) 'Using interaction data to explain difficulty navigating online', *ACM Transactions on the Web (TWEB)*, Vol. 8, No. 4, pp.1–24.

Vázquez, A., Oliveira, J.G., Dezsö, Z., Goh, K-I., Kondor, I. and Barabási, A-L. (2006) 'Modeling bursts and heavy tails in human dynamics', *Physical Review E*, Vol. 73, No. 3, p.036127.

Wise, A., Speer, J., Marbouti, F. and Hsiao, Y-T. (2013) 'Broadening the notion of participation in online discussions: examining patterns in learners' online listening behaviors', *Instructional Science*, Vol. 41, No. 2, pp.323–343.

Wise, A.F., Perera, N., Hsiao, Y-T., Speer, J. and Marbouti, F. (2012) 'Microanalytic case studies of individual participation patterns in an asynchronous online discussion in an undergraduate blended course', *The Internet and Higher Education*, Vol. 15, No. 2, pp.108–117.

Xing, W., Guo, R., Petakovic, E. and Goggins, S. (2015) 'Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory', *Computers in Human Behavior*, Vol. 47, pp.168–181.

Zahoor, S., Rajput, D., Bedekar, M. and Kosamkar, P. (2015) 'Capturing, understanding and interpreting user interactions with the browser as implicit interest indicators', *Proceedings of the International Conference on Pervasive Computing*, Istanbul, Turkey, pp.1–6.

Zaïane, O.R. and Luo, J. (2001) 'Web usage mining for a better web-based learning environment', *Advanced Technology for Education*, pp.60–64.