Towards Reading Session-based indicators in Educational Reading Analytics

Madjid Sadallah[1,2], Benoît Encelle[3], Azze-Eddine Maredj[2], and Yannick Prié[4]

[1] Computer Science Department, University of Bejaia, 06000 Bejaia, Algeria
[2] CERIST, Algiers, Algeria
{msadallah,amaredj}@cerist.dz,
[3] LIRIS - UMR 5205 CNRS, France
bencelle@liris.cnrs.fr,
[4] LINA - UMR 6241 CNRS, France,
yannick.prie@univ-nantes.fr

Abstract. It is a challenging task to identify eLearning courses parts that have to be revised to best suit learners' requirements. Reading being one of the most salient learning activities, one way of doing so is to study how learners consume courses. We intend to support course authors (e.g. teachers) during courses revision by providing them with reading indicators. We use the concept of reading session to denote a learner's active reading period, and we provide several associated reading indicators. In our server-side approach, reading sessions and indicators are calculated using web server logs. We evaluate the relevance of our proposals using logs from a major French eLearning platform. Results are promising: calculated reading sessions are theoretically more precise than other best applicable approaches, and course authors consider suggested indicators to be appropriate to courses revision. Using reading sessions and associated indicators could facilitate authors' work of course reengineering.

Keywords: Reading analytics, Reading monitoring, Reading indicators, Reading sessions, Web log mining

1 Introduction

More and more educational documents have been made available online by thousands of authors, and are being accessed by millions of learners each day, be it for rapid consultation or through active reading. Authors of online educational material can obtain automated feedback on how learners proceed through it. Such information may help them to get hints on users' learning experiences, spot reading issues or infer what knowledge is exactly gained from courses. This can help them make informed decisions on how their documents should be improved, from local clarifications to deeper restructurations. Our general interest is related with helping authors do the maintenance and evolution of eLearning documents. Educational settings are cyclic in nature and authors can take benefit of each cycle to evolve their courses to maximize their learning efficiency

(by being more precise, more comprehensible, more adapted to learners' needs, etc.). We investigate the use of learners' reading logs to support authors during these adaptation/improvement stages. To achieve this goal, we advocate "usage-based document reengineering" [28], a process defined as a kind of reengineering that changes document content and structures based on the analysis of readers' usages as recorded in logs.

In this paper, we focus on building indicators of learners' reading behaviors for authors to understand how their courses are used. We build on related work concerning reading-based indicators and session identification (Section 2). Using consumption logs of several courses provided by a major French eLearning platform (Section 3), we present our algorithm for computing *reading sessions* that denote active learners reading periods (Section 4). We evaluate the quality of our algorithm by comparing the results with other methods, as well as categories of reading indicators constructed from reading sessionsand we estimate the usefulness of our indicators through a survey (Section 5).

2 Related work

Monitoring reading. Reading is a fundamental activity and the basis of learning. Different actions and interactions between the learner and the reading material are at the core of reading. Studies have shown that capturing and interpreting this data are an effective means to reflect and predict, with good precision, the users' reading usages and behavior [12, 31]. Crawler-gathered data, originated from the server side data, the client side data or from both the sources, are often used. In education, a high level interpretation of this data is performed by assessing various computed metrics called indicators [9]. The community propositions range from simple measures and usage statistics like number of visits or visits per web page [24] to more advanced indicators like inferring students' attitudes that affect learning [1] and predicting students' knowledge [10].

Solely relying on request-based information to study reading has a major drawback that requesting a page is not necessarily equivalent to reading everything that is presented on it [12]. The use of time between requests has been shown to be an effective and unobtrusive way to improve the derived assumptions without interfering with user behavior and environment [14]. However, a difficulty arises from the fact that time generally cannot be directly obtained from the events stream web logs only contain actions timestamps with no explicit markers of their ends. Hence, each action duration must be estimated as the time difference between its occurrence and the subsequent one. Such a method could be imprecise since inactivity periods may be contained within the assessed duration. A recent study demonstrates that the adoption of a particular estimation strategy can have a significant impact on the fit of analytical models of learners' performance and their interpretation [18]. The precision of indicators highly depends on the correct estimation of action durations.

Sessions-based indicators. A body of work targeted the estimation of action durations using web log analysis (see [18] for an extensive background on time-

on-task estimation methods). Any adopted method aims to identify active and inactive periods from a user trail. The set of the actions performed by one user in a sustained and continuous activity is often referred to as a "session". A session can be seen as *a delimited and sustained set of pages visited by the same user within the duration of one particular visit to a particular website.*

The session-based indicators encode the navigation behavior of users over time [22], a valuable aspect that advocates their use to analyze reading efficiently beyond the course and page levels perspectives. Appropriately selecting these indicators have shown to provide good insightfulness in understanding learners activity. Learning sessions (defined in [20] as a set of sessions) duration and length were used to estimate the success and difficulty of the learning task [2]. Analysis of number of visits per session is studied in [16]. Navigation properties within them are studied in [4] and many navigation patterns were extracted to help evaluate and interpret online course activities [32] and to give insight into dependencies between page requests [17]. However, apart from such general session descriptive metrics common in web-based navigation, session-based reading indicators are not explicitly addressed. There is even no common definition of the "session" concept in eLearning.

Sessions identification. In eLearning, some authors use specific approaches for session identification like estimating the needed time for reading using an average reading time (in words per minute) [5] or learners self-reports on the time spent [26]. However, methods originated from the field of Web Usage Mining are often used. This field, which aims to reveal the knowledge hidden in navigation logs [23], has two main classes of approaches to deal with session identification problem: 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 [6], and if the duration of accumulated page view-times exceed 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 [7]. 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. The methods using time-based thresholds are the most commonly used in eLearning and are recommended in [20], without providing precise threshold values. As no generalized model exists for estimating a threshold in a given situation [15], authors rely on their data corpus and context characteristics to define thresholds values: 30 min [8], 60 min [30] or even 7 hours [25].

Sessions identification issues in eLearning. The time-based approach for identifying sessions best suits the needs of our study context; moreover, the navigation-based method cannot be used given that connections may exist between all course

webpages. However, we identified two main drawbacks of using a unique time threshold value (for pages or for sessions): 1) *eLearning activities are diverse*: 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 session time threshold [23]. As educational websites may have a complex structures and content, 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). Consequently, each part/page of the same course webpages is different (with regard with its inner-complexity) and thus requires a dedicated reading time.

3 Study objectives, data source and corpus description

This paper is concerned with solely reading activity and proposes to define for each course webpage the most suitable time needed by learner to achieve its reading. This time is used to specialize the concept of session in reading and to draw specific indicators beyond the aforementioned general ones. The detected learners' sessions of reading will allow not only to give more precision and expressiveness to existing reading indicators (e.g. durations, visits and revisits) but also to enrich reading analysis toolkit with specific session-based indicators.

Our proposals are implemented and evaluated using data from the major French e-learning platform *OpenClassrooms*[5]. Course authors are generally domain-experts, some of them are academic teachers and instructors. Our corpus data is constituted from web logs on 842 courses (mainly in computing and information technologies) over a 75 days period. Courses are organized as the nesting of *parts* (corresponding to chapters, sub-chapters, sections, etc.), each part being contained in a dedicated webpage. Logfiles contain information about website visitor activity and are automatically created by the web server. Common cleaning and preprocessing 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 part*.

Statistics about the corpus are presented in Table 1 (left) along different facets: number of requests, number of course parts, total number of distinct web-sessions and number of distinct authenticated readers. As the standard deviation (SD) values indicate, there's evidence of distribution inequality of these variables within the different courses. We selected a subset of 4 courses for our experiments as follows: 2 courses with the nearest values to respectively the me-

[5] With more than 850 courses and 1 million members, OpenClassrooms totalizes about 2.5 million unique visitors every month. See `http://www.openclassrooms.com/`

dian (`Screensaver`) and the mean (`Nodejs`), the most popular course (`Java`) and an atypical one (`XML`). Their respective properties are given in Table 1 (right).

Table 1. Statistics for (Left) all the 842 courses and (Right) the 4 selected courses

	Median	Mean	SD
Parts	17	28	35
Actions	8 922	45 681	197 549
Sessions	2 055	6 283	22 536
Users	199	476	1 252

	Screensaver	XML	Node.js	Java
Parts	11	107	36	164
Actions	4 380	10 284	61 387	1 099 295
Sessions	678	486	6 377	140 508
Users	80	34	611	4 582

4 Reading sessions

To study reading in eLearning, we use the concept of "*reading session*" to denote the period during which a reading activity takes place. It refers to a 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 characterize reading, for instance on Wikipedia [19]. We manage to identify reading sessions using course part (i.e. page) reading-time thresholds which are computed based on actual time spent by learners on parts. Since part thresholds are used to delimit reading sessions, we define a threshold as the maximum time needed for reading the corresponding part. Our method is composed of five consecutive steps, with the first two ones (*user identification* and *actions duration estimation*) as preprocessing steps. In the remainder of this paper, we will use *action* as a shortcut for *reading action*. The synthetic algorithm is given in Listing 1.

User identification. 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 proposal, 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.

Actions duration estimation. 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 are not directly available. As a consequence, 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.

Algorithm 1: Synthetic algorithm for reading session computation

```
// 1. Computing end timecodes and durations
foreach User in Data do
    foreach (Action,NextAction) of User do
        Action.End = NextAction.Begin
        Action.Duration = Action.End − Action.Begin
// 2. Computing parts thresholds
foreach Part in Data do
    PartData = Actions from Data, observed on Part
    Part.Threshold = Max(Peirce(PartData.durations))
// 3. Computing reading sessions per user
foreach User in Data do
    FirstAction = first Action of User
    FisrtAction.RS = 1
    foreach (Action,NextAction) of User do
        if Action.duration <= Part.Threshold then NextAction.RS =
        Action.RS;
        else NextAction.RS = Action.RS + 1;
```

Part-threshold values. Server-based monitoring can lead to very large durations, up to days, for parts that can be read in a couple of minutes. This is because a user may access a course part then change his activity momentously, for a long time or definitively. Moreover, some actions may be very short and hence not correspond to actual reading actions. To minimize the impact of these actions on the threshold calculation, we solely use "normal actions", excluding duration-excessive and duration-insignificant actions. This is performed by applying Peirce's criterion, a method that eliminates the presence of several suspicious data values (outliers) [27]. The maximum value of the remaining subset is then taken as the part reading threshold.

Dealing with unknown durations. Unknown durations occur for the last action since no other request can be used to define its end time. In order not to affect the corpus, and rather than skipping these actions, we assign them with the threshold values of the read parts.

Delimiting reading sessions. Using the reading thresholds, actions of each user are grouped into reading sessions. A reading session is assumed finished when the time spent for reading a part is greater than the time threshold of that part.

Example. Figure 1 presents data about the first three reading sessions of a randomly chosen reader of the `Node.js` course. The data include the count of the read parts, the cumulative total and unique read parts compared to the 36 parts of the course, ordered list of the read part, a graphical representation of the session and its total duration.

User	Reading session	parts count	Total read parts	Start part	Path	Path Graph	Duration
175	1	7	07/36	3	3;4;5;6;10;34;36		2m
	2	8	10/36	2	2;3;4;5;6;7;8;10		15m 2s
	3	4	13/36	10	10;11;12;13		6m 56s

Fig. 1. `Node.js` course: data about the reading sessions of a user

5 Reading session-based indicators

We have defined several reading indicators based on reading sessions, organized in four categories. Each category is supplemented with higher level reading indications, computed from its indicators, to denote important detected reading facts and possible issues. The whole set consists of 27 indicators and 21 reading indications [6]. These indicators make use of subsets or the whole of the reading sessions, they can be related to one or many courses and to one or many users. In the following, we present these categories with some indicators as illustrations.

5.1 Category 1: General facts about course reading

The distribution of the reading sessions over several dimensions can highlight many facts about learners' readings and provide basic hints for readings characterization within reading sessions. For instance, *Reading session duration* can serve as an indication of success and difficulty of the learning task (i.e. reading and understanding) [2]. Figure 2 presents three indicators on the four courses.

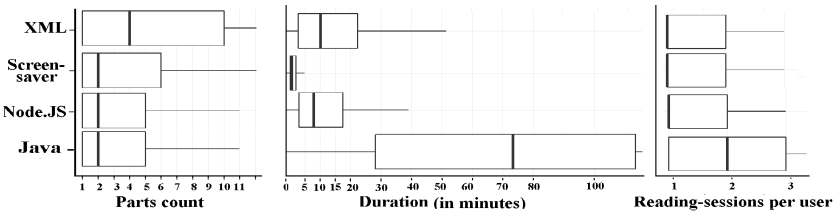

Fig. 2. Boxplots of three global indicators for the four courses, based respectively on reading sessions: number of parts, duration and number per user

5.2 Category 2: Reading paths and transitions

The reading path of a user is the sequence of parts that have been within his reading sessions. We analyse the most *Redundant/representative paths* and their *Deviation from the author expected one*. This allows to estimate *Global and per reading session progression ratio* and to detect *Paths covering the entire course.*

[6] The full set of indicators is briefly described at `http://bit.ly/reading-indicators`

Navigation properties were found to be correlated with learning task success [21]. For instance, the user path and its deviation from the optimal one are used to predict a possible user disorientation [11]. A user transition between two course parts allows reporting a relation between them. We have studied for each part the *Identities of the provenance and destination parts* and the *Linearity* of these transitions. This may help to point out needed restructuration of the course.

Figure 1 contains the reading path of a randomly chosen reader of the `Node.js` course. The traversal is mainly linear with some jumps especially in the first reading session. Some of the skipped parts have been read later in the following reading sessions (e.g.: parts 2 and 7). The linearity of the reading within the `XML` course is illustrated in Fig. 3 (left), together with provenance and destination for one specific part (right).

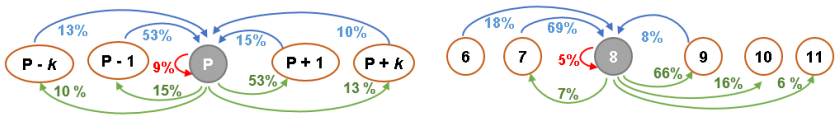

Fig. 3. Transitions ratios in `XML` reading: (Left) from and to a typical part, and (right) from and to *Part 8* ratios. In green: destination, in blue: provenance, in red: reread

5.3 Category 3: Rereading

Revisitation is a common browsing behavior and one of the most used strategies in reading for learning. It can indicate reading facts like potential users disorientation [13], parts popularity and detecting relationship between parts. We differentiate rereads that occur on the same reading session (*Within-session rereads*) from those that are performed on different reading sessions (*Between-session rereads*). Reading the same content many times is a potential indicator that readers are struggling with it. Between-session rereads may indicate that the reader needs a reminder of the earlier read parts (e.g.to understand new concepts presented later or to replace himself within the course context). Figure 4 presents the rereading data of the `Node.js` course. Ratio for the two reread types are presented on the left: parts 3 and 13 have an important within-session rereads ratio that may indicate their difficulty. Distribution of these rereads are presented on the right: rereads are mainly done within the same reading sessions, which may reflect potential difficulties in reading the course.

5.4 Category 4: Reading session interruption

Analyzing reading session interruptions allows determining the parts where reading stops definitively (*Final stop parts*) or with later resume (*Reading session break parts*). Resuming may take place on the same part (*Self resume*) or on

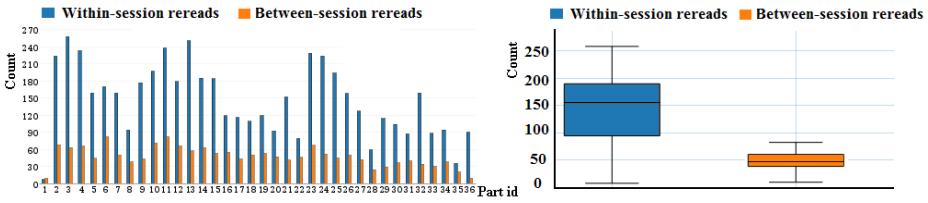

Fig. 4. Reread in `Node.js` course: (Left) Reread distribution on the first 20 parts (Right) Boxplot of the two reread types

the next one (*next resume*), cases which can be seen as normal. Resuming may also take place on distant forward or backward parts (*Back resume* and *Forward resume*) which may indicate the need for a more detailed analysis. Figure 5 illustrates these cases for the `Java` course. 82% of interruptions are final and concern parts 8 and 16. Resumes are often done on the same interruption parts and in some cases are back or forward. A deeper analysis may indicate whether there is an issue in the course structuring since resuming on the next part has the less important value or whether it is a normal reading pattern for this course.

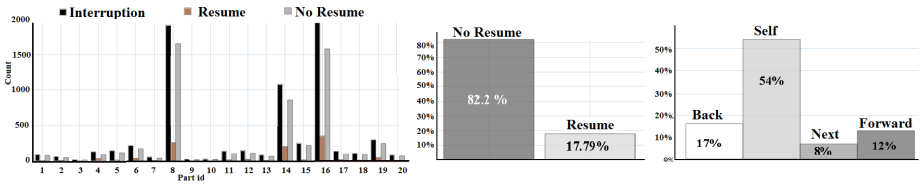

Fig. 5. Reading interruptions and resumes in `Java` course: (left) distribution on the course parts; (center) ratios of different types of interruption; (right) ratios of different types of resume

6 Evaluation and discussion

6.1 Reading session calculation

Quality of the reconstruction. As no effective measurement can assess the compliance of the reconstructed sessions with the actual ones, the reconstruction quality is often measured using the empirical observed Power Law distribution proposed in [3, 29]. This law states that most visits to a website are concentrated on a small number of pages. Evaluating this can be performed by a linear regression on the logarithm of the number of the distinct read parts 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. Results for our method applied on the four courses are shown on Fig. 6 as well as results using our method and the two other time-based methods with two threshold values: 10 min for pages and 30 min for sessions are shown on Table 2. They confirm our method capabilities in our context of study since it gives good fit results with an acceptable accuracy given by the error values.

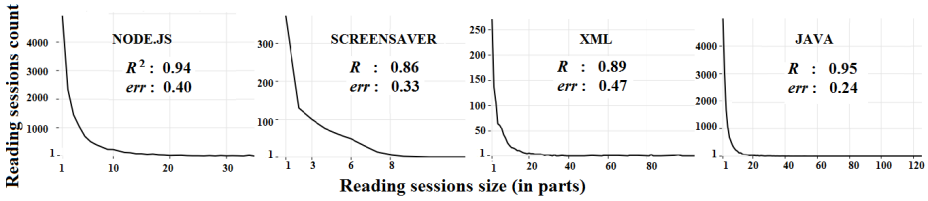

Fig. 6. Session size found by the power law distribution on the four courses

Table 2. Constructed sessions using three methods : our proposal, fixed page threshold (10-min) and fixed session threshold (30-min).

	Reading Session		10-min Page Thr.		30-min Session Thr.	
	R^2	*Err*	R^2	*Err*	R^2	*Err*
Nodejs	0.94	0.40	0.92	0.42	0.87	0.31
Screensaver	0.86	0.33	0.76	0.48	0.27	0.20
XML	0.89	0.47	0.82	0.45	0.79	0.51
Java	0.95	0.24	0.94	0.23	0.94	0.25

Compliance with parts size and complexity. We estimated the size of each part 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 part size (computed as the words and figure count) and time threshold for that part 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 part size without needing to calculate it for each part. We can make the hypothesis that it is also the case for part complexity, even if part size does not directly indicates complexity level of the content. This also confirms the need to take into account characteristics of the parts for more accurate threshold values.

Comparison with fixed page threshold method. Following a per page reading time threshold approach for delimiting reading sessions based on learners reflects the actual usages and differentiates parts based on their content. Using a sole fix value would give imprecise results since we assume for each part the same maximum reading time. In fact, some parts may be read faster while others

may need more than the fixed value time for reading, as exemplified by the four courses on Fig. 7. Whatever the selected threshold, there will always be parts which can be read in less than this time and others that may take more time.

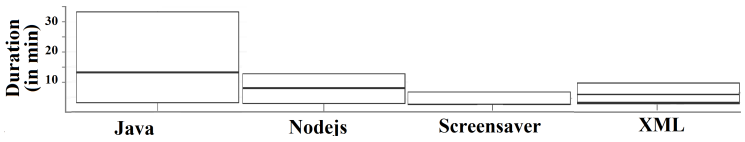

Fig. 7. Parts reading duration (boxplots): all are under 10 min for Screensaver.

Comparison with fixed session threshold method Our method does not define constraints on the session size, which would cut many continuous long sessions and merge other short ones. We used a 30 minutes threshold on the data of about 2000 distinct readers of the `Java` course, which is known to be quite complex and having long parts (and hence, time-demanding). The results gave a median of one (1) part – which mean that the half of sessions does not exceed one part – and a value of the 3rd quartile of 3 parts. The results using our method seem more realistic since they gave a median of 3 parts and 3rd quartile of 7 parts.

6.2 Reading session indicators apprehension by authors

To evaluate the relevance of the proposed indicators for course rewriting, we have conducted an exploratory study to gather OpenClassrooms course authors' opinions regarding our proposals. An online survey was set up that both contains *Likert scales* for rating indicators usefulness (very useful, useful, no opinion, somewhat useful or not useful) and free comment sections. At the time of writing, 105 out of hundreds authors had filled the online survey [7].

Indicators rating. Figure 8 presents results of authors' ratings aggregated by category. All the proposed indicators were highly rated (useful), with the more contrasted results on the indicators related to parts distribution within sessions (*average number of parts per session*), rereading (*rereaders count* and *rereading – within and between-session – distribution on the parts* and interruption points (*definitive stops parts*). Indicators related to the *average reading speed per part* and *session count on the course* are the least rated ones.

Comments and opinions. The authors acknowledge that the exchanges between authors and readers are essential to build interesting and productive courses. The fact that readers' usages logs allow to consider the end-user perspective on consuming the course is deemed interesting. All the authors assess the usefulness

[7] The questionnaire and full results are available at `http://bit.ly/authors-survey`

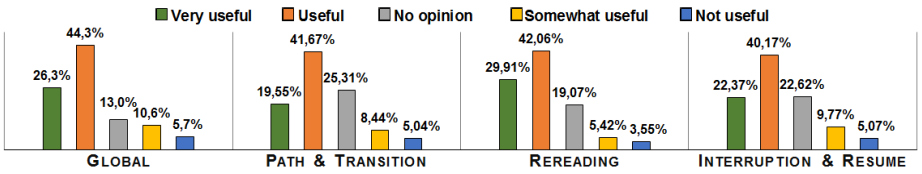

Fig. 8. Authors rating for the indicators, aggregated by categories

of reading usages to detect parts or aspects of the course that necessitate review. Many authors have appreciated indicators calculated by aggregating data, which may better reflect recurrent reading problems. They found these indicators judicious to give a good idea of the way learners read courses. While more than 73 authors estimate the set of indicators comprehensive enough to analyze 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. Another aspect reported by three authors is related to privacy; they suggested asking learners before logging them. Several authors proposed to consider the supplementation of computed indicators with explicit readers' feedback (courses and parts ratings, comments and annotations, etc.) that would help them to better understand readers' needs.

6.3 Conclusion and future work

In this paper, we use the concept of reading session as a means to model course reading. The proposed approach for session identification is grounded on data that represent learners' interactions with course parts and take into account each part characteristics. The computed threshold values for part reading-times are dynamic since they may be updated when new reading actions are logged. This allows their automatic adjustment 1/to precise their values with incoming reading data and 2/to take into account any evolution of the courses like pages restructuring and content update. Consequently, this approach seems a very plausible way to simulate learners reading and to fit the expected statistical behavior of real reading sessions. We plan to further verify this method capabilities by using/ defining appropriate metrics to characterize parts complexity and to compare the deduced sessions compliance with the real ones.

Several reading indicators based on reading sessions have been proposed and illustrated on courses from a major French eLearning provider. Reading session-based indicators intend to analyze reading from a behavioral perspective, a viewpoint that may efficiently reflect potential readers' needs and reading issues. This is acknowledged by authors whose survey responses indicate 1/ their understanding of the proposed set of indicators and 2/ the potential relevance of

these indicators for course revision. Most of the gathered comments and suggestions actually 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, supplement these with readers' annotations and build a simple and intuitive dashboard to assist authors.

References

1. I. Arroyo, T. Murray, B. P. Woolf, and C. Beal. Inferring unobservable learning variables from students help seeking behavior. In *Proceedings of International Conference on Intelligent tutoring systems*, pages 782–784. Springer, 2004.
2. A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 35–44, New York, NY, USA, 2010. ACM.
3. B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. In *Workshop on Web Mining*, pages 7–14, 2001.
4. N. Bousbia, I. Rebaï, J.-M. Labat, and A. Balla. Learners' navigation behavior identification based on trace analysis. *User Modeling and User-Adapted Interaction*, 20(5):455–494, 2010.
5. A. H. Brown and T. Green. Time students spend reading threaded discussions in online graduate courses requiring asynchronous participation. *The International Review of Research in Open and Distributed Learning*, 10(6):51–64, 2009.
6. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, pages 558–567. IEEE, 1997.
7. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, 1(1):5–32, 1999.
8. R. del Valle and T. Duffy. Online learning: Learner characteristics and their approaches to managing learning. *Instructional Science*, 37(2):129–149, 2009.
9. A. Dimitrakopoulou, A. Petrou, A. Martinez, J. A. Marcos, V. Kollias, P. Jermann, A. Harrer, Y. Dimitriadis, and L. Bollen. State of the art of interaction analysis for metacognitive support & diagnosis. 2006.
10. M. Feng, N. T. Heffernan, and K. R. Koedinger. Looking for sources of error in predicting student's knowledge. In *Proceedings of the 2005 AAAI Workshop on Educational Data Mining*, pages 54–61, 2005.
11. J. Gwizdka and I. Spence. Implicit measures of lostness and success in web navigation. *Interacting with Computers*, 19(3):357–369, 2007.
12. D. Hauger, A. Paramythis, and S. Weibelzahl. Using browser interaction data to determine page reading behavior. In *User Modeling, Adaption and Personalization*, pages 147–158. Springer, 2011.
13. E. Herder. Revisitation patterns and disorientation. In *11. GI-Workshop" Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen"*, 2003.
14. K. Hofmann, C. Reed, and H. Holz. Unobtrusive data collection for web-based social navigation. In *Workshop on the Social Navigation and Community based Adaptation Technologies*. Citeseer, 2006.
15. T. Huynh and J. Miller. Empirical observations on the session timeout threshold. *Inf. Process. Manage.*, 45(5):513–528, Sept. 2009.

16. I. Kazanidis, S. Valsamidis, S. Kontogiannis, and A. Karakos. Measuring and mining lms data. In *Informatics (PCI), 2012 16th Panhellenic Conference on*, pages 296–301, Oct 2012.
17. A. Klašnja-Milićević, B. Vesin, M. Ivanović, and Z. Budimac. E-learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, 56(3):885–899, 2011.
18. V. Kovanovic, D. Gaševic, S. Dawson, S. Joksimovic, R. S. Baker, and M. Hatala. Penetrating the black box of time-on-task estimation. pages 184–193, New York, New York, USA, 2015. ACM Press.
19. J. Lehmann, C. Müller-Birn, D. Laniado, M. Lalmas, and A. Kaltenbrunner. Reader preferences and behavior on wikipedia. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 88–97. ACM, 2014.
20. C. G. Marquardt, K. Becker, and D. D. Ruiz. A pre-processing tool for web usage mining in the distance education domain. In *Database Engineering and Applications Symposium, 2004. IDEAS'04*, pages 78–87. IEEE, 2004.
21. J. E. McEneaney. Graphic and numerical methods to assess navigation in hypertext. *International Journal of Human-Computer Studies*, 55(5):761–786, 2001.
22. B. Mobasher. Data mining for web personalization. In *The adaptive web*, pages 90–135. Springer, 2007.
23. M. Munk and M. Drlik. Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. In *Digital Information and Communication Technology and Its Applications*, pages 60–74. Springer, 2011.
24. C. Pahl and D. Donnellan. Data mining technology for the evaluation of web-based teaching and learning systems. 2002.
25. D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(6):759–772, 2009.
26. M. Romero and E. Barberà. Quality of e-learners' time and learning performance beyond quantitative time-on-task. *The International Review of Research in Open and Distributed Learning*, 12(5):125–137, 2011.
27. S. M. Ross. Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, 20(2):38–41, 2003.
28. M. Sadallah, B. Encelle, A.-E. Maredj, and Y. Prié. A framework for usage-based document reengineering. In *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13, pages 99–102, New York, NY, USA, 2013. ACM.
29. A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.
30. A. Wise, J. Speer, F. Marbouti, and Y.-T. Hsiao. Broadening the notion of participation in online discussions: examining patterns in learners' online listening behaviors. *Instructional Science*, 41(2):323–343, 2013.
31. S. Zahoor, D. Rajput, M. Bedekar, and P. Kosamkar. Capturing, understanding and interpreting user interactions with the browser as implicit interest indicators. In *Proceedings of the 2015 International Conference on Pervasive Computing (ICPC)*, pages 1–6. IEEE, 2015.
32. O. R. Zaíane. Building a recommender agent for e-learning systems. In *Proceedings of International Conference on Computers in Education*, pages 55–59. IEEE, 2002.