Disengagement Detection in On-line Learning: Validation Studies and Perspectives

Mihaela Cocea and Stephan Weibelzahl

Abstract— Learning environments aim to deliver efficacious instruction, but rarely take into consideration the motivational factors involved in the learning process. However, motivational aspects like engagement play an important role in effective learning–engaged learners gain more. E-Learning systems could be improved by tracking students' disengagement that, in turn, would allow interventions at appropriate times. This idea has been exploited several times for Intelligent Tutoring Systems, but not yet in other types of learning environments that are less structured. To address this gap, our research looks at on-line content-delivery systems using educational data mining techniques. Previously, several attributes relevant for disengagement prediction were identified by means of log-file analysis on HTML-Tutor, a web-based learning environment. In this paper, we investigate the extendibility of our approach to other systems by studying the relevance of these attributes for predicting disengagement in a different e-Learning system. To this end, two validation studies were conducted indicating that the previously identified attributes are pertinent for disengagement prediction, and that two new meta-attributes derived from log data observations improve prediction and may potentially be used for automatic log-file annotation.

Index Terms— e-Learning, educational data mining, disengagement prediction, log-file analysis

1 Introduction

EDUCATIONAL software strives to meet the learners' needs and preferences in order to make learning more efficient; the complexity is considerable and many aspects are taken into consideration. However, most systems do not consider the learner's motivation for tailoring teaching strategies and content, despite of its great impact on learning being generally acknowledged and of research showing that lack of motivation is correlated with learning rate decrease (e.g. [1]).

A number of attempts have been undertaken to accommodate the learner's motivational states, mostly by means of design. E-Learning systems attempted to motivate students through an attractive design by using multimedia materials or by including game features that have great potential [2] and have been proved successful in a number of cases (e.g. [3]). Despite these efforts, students are not always focused on learning and even try to game the systems (attempting to succeed in an educational environment by exploiting properties of the system's help and feedback rather than by attempting to learn the material)[1] .

Learner's self-assessment has been used for a long time in classroom context, and recently also in e-Learning, where it has been proved to be reliable, and a valuable and accurate source of motivational information [4].

However, to effectively address the motivational factors that influence learning they need to be assessed for each individual to allow personalized interventions based on this assessment. To do this efficiently, automatic analysis is necessary.

The learner's actions preserved in log files have been relatively recently discovered as a valuable source of information and several approaches to motivation detection and intervention have used log-file analysis. An important advantage of log file analysis over self-assessment approaches is the unobtrusiveness of the assessment process – that would be similar to the classroom situation where a teacher observes that a learner is not motivated without interrupting his/her activities.

Several efforts to detect motivational aspects from learners' actions are reported in the literature [5], [1], [6], [7], [8], [9], [10], [11]. However, all these efforts are concentrated on Intelligent Tutoring Systems or problem-solving environments. As on-line content-delivery systems are increasingly used in formal education, there is a need to extend this research to encompass this type of systems as well. The interaction in these systems is less constrained and structured compared with problem-solving environments, posing several difficulties to an automatic analysis of learners' activity. To address this challenge, we restricted our research to one motivational aspect, disengagement, and looked at identifying the relevant information from learners' actions to be used for its prediction.

Using data from a web-based interactive environment, HTML-Tutor, we have identified six relevant attributes by means of educational data mining techniques [12]. In

- *M. Cocea is with the London Knowledge Lab, Birkbeck College, Department of Computer Science, 23-29 Emerald Street, London, WC1N 3QS E-mail: mihaela@ dcs.bbk.ac.uk*
- *S. Weibelzahl is with the National College of Ireland, School of Computing, Mayor Street, Dublin 1. E-mail: sweibelzahl@ ncirl.ie*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

this paper we investigate the extendibility of our approach to other systems by studying the relevance of these attributes for predicting disengagement in a different e-Learning system.

The rest of the paper is structured as follows. In Section 2 previous work related to motivation and engagement prediction is presented. Section 3 briefly presents the log-file analysis performed on HTML-Tutor data by which the relevant attributes for disengagement prediction were identified. Section 4 includes the two validation studies conducted on iHelp data and Section 5 discusses the results and implications of the validation studies, and relates our outcomes with the previous approaches to engagement prediction. Section 6 discusses several perspectives on the outcomes of this research and its possible impact, and concludes the paper.

2 Related research

Before presenting related research on detection of motivational aspects, a brief outline is given on how engagement is related to other motivational concepts.

Motivational research [13] makes used of several concepts, besides motivation itself: engagement, interest, effort, focus of attention, self-efficacy, confidence etc. The research presented in this paper focuses on engagement, or rather on disengagement, as an undesirable motivation state. For our purposes, a student is considered to be engaged if he/she is focused on the current learning activity and disengaged otherwise. A number of concepts in motivational research such as interest, effort, focus of attention and motivation are related, though not identical, to engagement (see e.g., [13]):

1. Engagement can be influenced by interest, as people tend to be more engaged in activities they are interested in; thus, interest is a determinant of engagement.
2. Effort is closely related to interest in the same way: more effort is invested if the person has interest in the activity. The relation between engagement and effort can be resumed by: engagement can be present with or without effort; if the activity is pleasant (and/or easy), engagement is possible without effort; in the case of more unpleasant (and/or difficult) activities, effort may be required to stay engaged.
3. The difference between engagement and focus of attention, as used in research, is that focus of attention refers to attention through a specific sensorial channel (e.g. visual focus), while engagement refers to the entire mental activity (involving at the same time perception, attention, reasoning, volition and emotions).
4. In relation to motivation, engagement is just one aspect indicating that, for a reason or another, the person is motivated to do the activity he/she is engaged in, or, on the contrary, if the person is disengaged, that he/she may not motivated to do the activity. In other words, engagement is an indicator of motivation.

Although there are several approaches to motivational issues in e-Learning, we restrict our review to only some of them that are related to detection of motivational aspects in general and engagement in particular, by means of using learners' actions.

Several approaches for motivation detection from learner's interactions with the e-Learning system have been proposed ranging from rule-based approaches to Bayesian networks.

A rule-based approach based on ARCS (Attention, Relevance, Confidence and Satisfaction) Model [14] has been developed [5] to infer motivational states from the learners' behavior using a ten-question quiz. A number of 85 inference rules were produced by the participants who had access to replays of the learners' interactions with the system and to the learners' motivational traits.

Another approach [8] based on ARCS Model is used to infer three aspects of motivation: confidence, confusion and effort, from the learner's focus of attention and inputs related to learners' actions: time to perform the task, time to read the paragraph related to the task, the time for the learner to decide how to perform the task, the time when the learner starts/finishes the task, the number of tasks the learner has finished with respect to the current plan (progress), the number of unexpected tasks performed by the learner which are not included in the current learning plan and the number of questions asking for help.

Engagement tracing [6] is an approach based on Item Response Theory that proposes the estimation of the probability of a correct response given a specific response time for modeling disengagement; two methods of generating responses are assumed: blindly guess when the student is disengaged and an answer with a certain probability of being correct when the student is engaged. The model also takes into account individual differences in reading speed and level of knowledge.

A dynamic mixture model combining a hidden Markov model with Item Response Theory was proposed in [9]. The dynamic mixture model takes into account: student proficiency, motivation, evidence of motivation, and a student's response to a problem. The motivation variable can have three values: a) motivated, b) unmotivated and exhausting all the hints in order to reach the final one that gives the correct answer: unmotivated-hint and c) unmotivated and quickly guessing answers to find the correct answer: unmotivated-guess.

A Bayesian Network has been developed [7] from log-data in order to infer variables related to learning and attitudes toward the tutor and the system. The log-data registered variables like problem-solving time, mistakes and help requests.

A latent response model [1] was proposed for identifying the students that game the system. Using a pretest–posttest approach, the gaming behavior was classified in two categories: a) with no impact on learning and b) with decrease in learning gain. The variables used in the model were: student's actions and probabilistic information about the student's prior skills.

The same problem of gaming behavior was addressed in [10], an approach that combines classroom observations with logged actions in order to detect gaming behavior manifested by guessing and checking or hint/ help abuse. Prevention strategies have been proposed [11]: two active interventions for the two types of gaming behavior and a passive intervention. When a student was detected to manifest one of the two gaming behaviors, a message was displayed to the student encouraging him/her to try harder, ask the teacher for help or pursue other suitable actions. The passive intervention had no triggering mechanism and consisted in providing visual feedback on student's actions and progress that was continuously displayed on screen and available for viewing by the student and teacher.

3. Engagement Prediction From Log Files

Our approach is different from the previous ones in the fact that it envisages prediction of engagement from both main activities encountered in e-Learning systems: reading and problem-solving activities. The two models based on IRT presented in the previous section work very well for problem-solving activities only, but they have the disadvantage of considering engagement after the learning activity. Tracking engagement when the student is learning (reading pages) allows intervention at appropriate times and before the self-evaluation of learning (problem solving), when bad performance could be caused by disengagement in answering the questions, but also by disengagement during learning time.

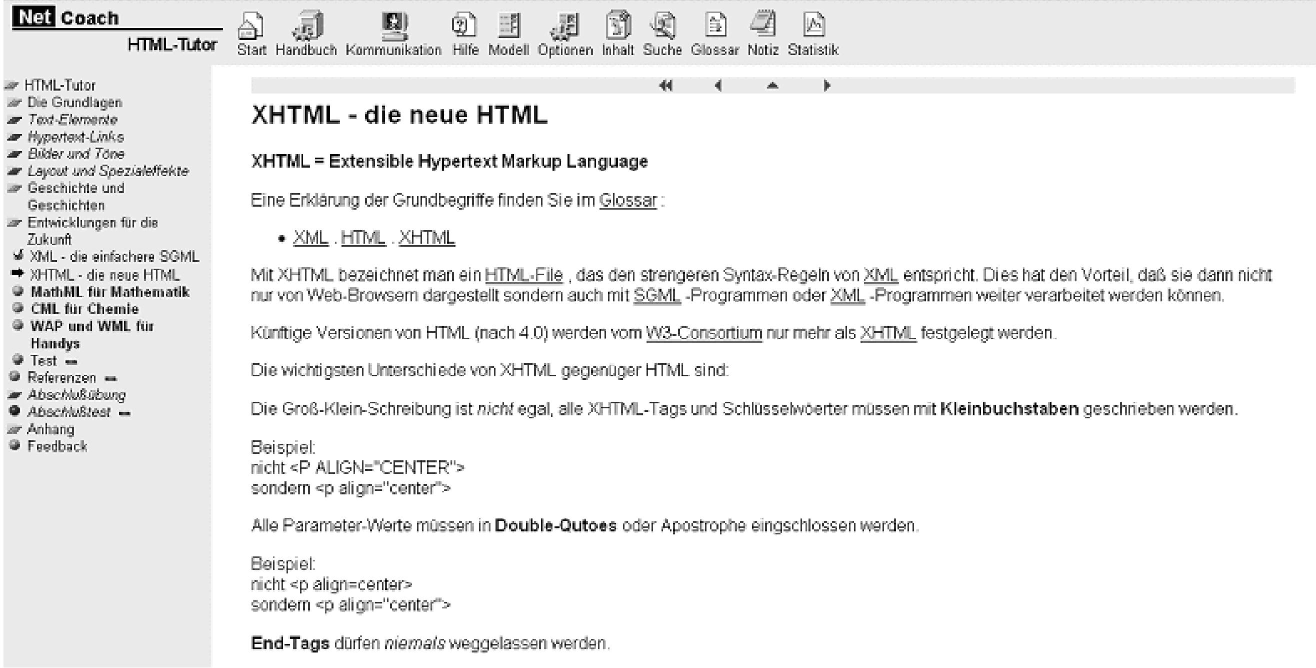
In previous research [12] we proposed a different approach to engagement prediction that would cover both learning and problem-solving activities typically encountered in e-Learning system.

We analyzed log files from HTML-Tutor – a web based interactive learning environment based on NetCoach [15]. The system is in German and is organized in seven high-level topics on HTML, e.g., hyperlinks, layout, XML, etc. In the screenshot displayed in Fig. 1, these topics are listed in the left side of the screen. Each high-level topic includes several sub-topics that may contain one or more items. Each component of this hierarchy links to a file that is displayed in the central area of the screen. A navigation bar is also present at the top of this central area. The top of the screen includes a toolbar with several icons linking to: a manual on how to use the system, communication tools, frequently asked questions, preferences on the display of information on the screen, a glossary, a notes tool and statistics tool about the personal usage of the system (e.g. coverage of topics, performance on tests).

The purpose of the analysis on the HTML-Tutor log data was twofold: (a) to identify attributes that are relevant for prediction and (b) to explore several prediction methods, mainly as a consistency check and secondly as a way to identify a best performing method (should it be the case). Consequently, three datasets were used to control the contribution of attributes and eight prediction methods were employed for the consistency of prediction.

Log files of 48 users were collected. These users spent between one and seven sessions, where a session is marked by login and logout. A pilot study [16] revealed that using sessions as units of analysis leaves no time for intervention as disengagement could be detected only after forty-five minutes of activity and most disengaged students would log out before that time. To overcome this problem, session were divided in sequences of ten minutes. From this process 1015 sequences were obtained: 943 sequences of exactly ten minutes and 72 sequences varying between 7 and 592 seconds.

From the fourteen logged events, a total of thirty attributes were derived. Two events – reading pages and taking tests – occurred considerably more often than all the others, with a frequency of occurrence of 850 and 458,



XHTML - die neue HTML

XHTML = Extensible Hypertext Markup Language

Eine Erklärung der Grundbegriffe finden Sie im Glossar :

- XML . HTML . XHTML

Mit XHTML bezeichnet man ein HTML-File , das den strengeren Syntax-Regeln von XML entspricht. Dies hat den Vorteil, daß sie dann nicht nur von Web-Browsern dargestellt sondern auch mit SGML -Programmen oder XML -Programmen weiter verarbeitet werden können.

Künftige Versionen von HTML (nach 4.0) werden vom W3-Consortium nur mehr als XHTML festgelegt werden.

Die wichtigsten Unterschiede von XHTML gegenüger HTML sind:

Die Groß-Klein-Schreibung ist *nicht* egal, alle XHTML-Tags und Schlüsselwöerter müssen mit **Kleinbuchstaben** geschrieben werden.

Beispiel:
nicht <P ALIGN="CENTER">
sondern <p align="center">

Alle Parameter-Werte müssen in **Double-Qutoes** oder Apostrophe eingschlossen werden.

Beispiel:
nicht <p align=center>
sondern <p align="center">

End-Tags dürfen *niemals* weggelassen werden.

Fig. 1. Screenshot of HTML-Tutor from XHTML Topic.

respectively, out of a total of 1015 sequences. Two other events - hyperlinks and glossary - were noticeably more frequent than the rest, with a frequency of 245 and 76, respectively, while the remaining ten events were rare (with an average of 16 occurrences in 1015 sequences). A few examples of these less frequent events are preferences, search and statistics. For a complete list of frequencies of all events, see [12] .

Based on the frequency of events, three datasets were defined: one that included attributes of all events, one that included the attributes of the four most frequent events and one that included only the two most frequent events. By doing this, we aimed to identify the relevant features, taking into consideration the sparsity of data at the same time.

Eight methods that were applicable to our data were employed [17], [18]:

1. Bayesian Nets with K2 algorithm and maximum 3 parent nodes (BN).
2. Logistic regression (LR).
3. Simple logistic classification (SL).
4. Instance based classification with IBk algorithm (IBk).
5. Attribute Selected Classification using J48 classifier and Best First search (ASC).
6. Bagging using REP (reduced-error pruning) tree classifier (B).
7. Classification via Regression (CvR).
8. Decision Trees with J48 classifier based on Quilan's C4.5 algorithm [21] (DT).

We describe here only the attributes of the two most frequent events, i.e. accessing pages and taking tests: number of pages, average time spent of pages, number of tests, average time spent on tests, number of correctly answered tests and number of incorrectly answered tests. The attributes of the other events are similar, typically including the frequency of access and the average time. For a complete list of attributes, see [12].

Each sequence was labeled as engaged, disengaged or neutral. Three human experts (designated as raters) were involved: rater 1 labeled all sequences, while rater 2 and 3 participated in a coding reliability study (more details in [6]). All raters used the unprocessed log files divided in sequences of 10 minutes containing all events. The output of the reliability study was a 92% agreement between raters, a Cohen's kappa [19] measurement of agreement of 0.83 ($p < 0.01$) and a Krippendorff's alpha [20] of 0.84, suggesting the annotation of sequences was conducted in a reliable fashion [21].

The results showed small variation of prediction values across methods and between the three datasets. Two indicators were especially considered: accuracy (the percentage of correct predictions), as an indication of the quality of prediction across all classes (engaged, disengaged, neutral) and true positive (TP) rate for the disengaged class as an indication of the extent of correct identification of disengaged learners. To give a complete picture and a grasp of the real meaning of the data, other indicators are included: the false positives (FP) rate for disengaged class, the precision indicator (TP/ (TP + FP)) for disengaged class and the mean absolute error. In our context, TP rate is more important than precision because it indicates the correct percentage from actual instances of a class, while precision indicates the correct percentage from predicted instances of that class.

Waikato Environment for Knowledge Analysis (WEKA) [17] was used to perform the analysis. Only sequences of exactly 10 minutes were used and from the 943 entries, 679 (72%) were used for training and 264 (28%) for testing.

Across methods, the prediction values varied between 84.85% (using IBk on third dataset) and 92.80% (using CvR on first dataset) accuracy. The variation of the true positive rate for the disengaged class was even smaller: between 0.91 and 0.96 (across all datasets and methods). Using the average across methods, the three datasets were compared: the first dataset performed best, with an average of 0.90% better accuracy than the second dataset and an average of 1.38% better than the third dataset; the second dataset performed better than the third dataset by 0.48%. The average variation of the true positive rate across datasets was negligible – less than 0.005. Given these relatively small variations and taking into consideration factors like sparsity of data and computational complexity, the attributes of the smallest dataset were considered the most relevant for a prediction model of disengagement. The results of the experiments for the smallest dataset are presented in Table 1.

TABLE 1

HTML-TUTOR: EXPERIMENT RESULTS SUMMARY

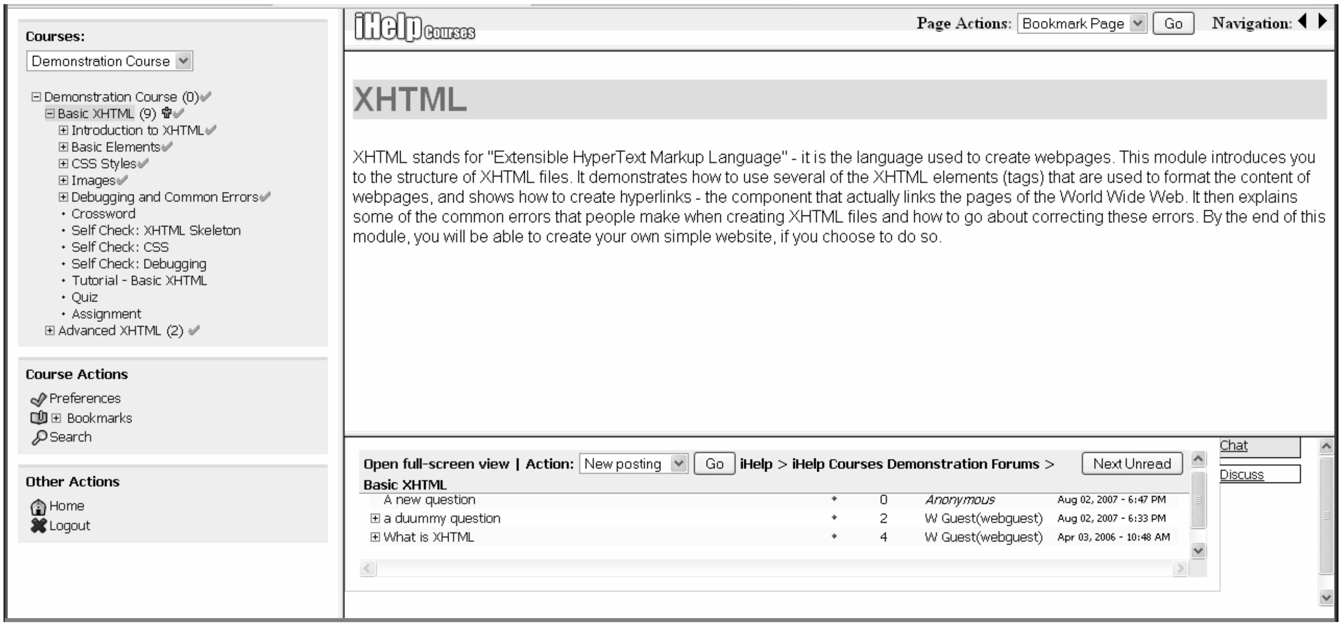
	BN	LR	SL	IBk	ASC	B	CvR	DT
%correct	89.77	86.36	87.12	84.85	90.91	89.77	90.91	90.15
TP rate	0.94	0.94	0.95	0.93	0.94	0.93	0.94	0.93
FP rate	0.18	0.29	0.28	0.24	0.16	0.17	0.16	0.16
Precision	0.91	0.86	0.87	0.89	0.92	0.92	0.92	0.92
Error	0.12	0.14	0.14	0.14	0.11	0.13	0.12	0.11

To summarize, relevant attributes for disengagement prediction were identified for HTML-Tutor. No method significantly outperformed the others, indicating consistency of prediction and allowing several possibilities for usage of the prediction methods as discussed in Section 5.

The next step was to investigate whether this approach worked on a different system and, more specifically, if the attributes identifies as being relevant for HTML-Tutor would be relevant for another system and, therefore, produce acceptable levels of prediction. A validation study was conducted for this purpose which is presented in the next section.

4 VALIDATION STUDIES

In order to validate our approach for engagement prediction presented above we analyzed data from iHelp, a web-based learning system developed and deployed at the University of Saskatchewan. This system includes two web based applications designed to support both learners



XHTML stands for "Extensible HyperText Markup Language" - it is the language used to create webpages. This module introduces you to the structure of XHTML files. It demonstrates how to use several of the XHTML elements (tags) that are used to format the content of webpages, and shows how to create hyperlinks - the component that actually links the pages of the World Wide Web. It then explains some of the common errors that people make when creating XHTML files and how to go about correcting these errors. By the end of this module, you will be able to create your own simple website, if you choose to do so.

A new question	•	0	Anonymous	Aug 02, 2007 - 6:47 PM
⊞ a duummy question	•	2	W Guest(webguest)	Aug 02, 2007 - 6:33 PM
⊞ What is XHTML	•	4	W Guest(webguest)	Apr 03, 2006 - 10:48 AM

Fig. 2. Screenshot of iHelp on XHTML content.

and instructors throughout the learning process: the iHelp Discussion system and iHelp Learning Content Management System (also called iHelp Courses). The former allows communication among students and instructors, while the latter is designed to deliver online courses to students working at a distance, providing course content (text and multimedia) and quizzes. The content is organized in packages that contain hierarchical activities. A single package is displayed at one time of the left of the screen, as illustrated in Fig. 2. Besides the structure of the package, on the left there are two menus, one related to *course actions*, such as preferences or search, and one related to other actions, such as logout. Each activity from the package is linked to a file that is displayed in the main area of the screen. At the top of this area, a navigation bar allows moving back and forward. Collaboration tools – chat and discussion forum – are available in the lower part of the screen.

The same type of data about the interactions was selected from registered information to perform the same type of analysis as the one performed on HTML-Tutor data. An HTML course was chosen to control the domain variable and, therefore, prevent differences in results caused by differences in subject matter.

Two studies were conducted with iHelp data. In the first study logged data from 11 users was used, meaning a total of 108 sessions and 450 sequences (341 of exactly 10 minutes and 109 less than 10 minutes). The second study included logged data from 21 students (all the students studying that course), meaning a total of 218 sessions and 735 sequences (513 of exactly 10 minutes and 222 less than 10 minutes).

4.1 Study 1

In the analysis several attributes mainly related to reading pages and quizzes events were used. These attributes are presented in Table 2. The terms tests and quizzes will be used interchangeably; they refer to the same type of problem-solving activity, except that in HTML-Tutor they are called tests and in iHelp they are named quizzes.

Given the smaller number of instances, sequences of less than 10 minutes were included in the analysis to see if the number of instances has an influence on prediction. As a consequence, to distinguish between these sequences and the ones of exactly 10 minutes, the total time of a sequence was included as an attribute. Compared to the analysis of HTML-Tutor logs, in the first study, for iHelp there are fewer attributes related to quizzes. Thus, information on number of questions attempted and on time spent on them is included, but information about the correctness or incorrectness of answers given by users was not available at the time of data retrieval.

For each 10 minutes sequence, the level of engagement was rated by an expert using the same approach as for HTML-Tutor that was briefly presented in Section 3. With HTML-Tutor, three level of engagement were used: engaged, disengaged and neutral. Neutral was used for situations when raters found it hard to decide whether the user was engaged or disengaged. With iHelp, this difficulty was not encountered. The rating consistency was verified on HTML-Tutor data by measuring inter-

TABLE 2

THE ATTRIBUTES USED FOR ANALYSIS

Code	Attribute description
NoPages	Number of pages read
AvgTimeP	Average time spent reading
NoQuestions	Number of questions from quizzes
AvgTimeQ	Average time spent on quizzes
Total time	Total time of a sequence

coding reliability. However, with iHelp only one rater classified the level of engagement for all sequences.

Two datasets were used in the analysis: DS1_S1 that included all sequences and DS2_S1 that included only sequences of exactly 10 minutes (S1 stands for Study 1). The same environment, WEKA, and the same eight methods were used for analysis; 10-fold cross-validation iterated 10 times was employed. The results are displayed in Table 3.

TABLE 3
STUDY 1: EXPERIMENT RESULTS

Dataset	Results	BN	LR	SL	IBk	ASC	B	CvR	DT
DS1_S1	%correct	81.73	83.82	83.58	84.00	84.38	85.11	85.33	84.38
	TP rate	0.78	0.82	0.81	0.79	0.77	0.79	0.80	0.78
	FP rate	0.14	0.14	0.14	0.11	0.08	0.08	0.09	0.08
	Precision	0.86	0.86	0.86	0.89	0.91	0.91	0.91	0.91
	Error	0.22	0.24	0.26	0.20	0.25	0.23	0.23	0.25
DS2_S1	%correct	84.29	85.82	85.47	84.91	84.97	85.38	85.26	85.24
	TP rate	0.78	0.77	0.76	0.77	0.75	0.76	0.75	0.75
	FP rate	0.10	0.06	0.06	0.08	0.06	0.06	0.06	0.06
	Precision	0.88	0.92	0.92	0.89	0.92	0.92	0.92	0.92
	Error	0.18	0.22	0.23	0.20	0.25	0.23	0.24	0.24

Compared to the results obtained on HTML-Tutor data, the prediction values are lower, especially for the true positive rates. However, the overall prediction is accurate on average more than 84% of the time and disengagement is still predicted correctly on average more than 78% of the time. Therefore, we can conclude the attributes used for prediction are relevant for iHelp as well.

Two differences between HTML-Tutor data and iHelp data may account for the lower true positive rates on the later: the smaller number of instances and the missing information about the correctness of answers on quizzes. To investigate their influence another study was needed.

During the labeling process of the iHelp data, a similarity was noticed with HTML-Tutor data in the patterns that disengaged students seemed to follow. Thus, some disengaged students spent a long time on the same page or test, while other students browsed very fast through content seemingly without reading. Based on these observations, we decided to include two attributes that reflected these aspects and investigate their potential role for an improved prediction.

Therefore, a second study was conducted to address the previously mentioned aspects – the role of more data, of data on performance on quizzes and of the two new attributes. The next section described this study and its results.

4.2 Study 2

To address the issue related to the number of instances, more data was processed and labeled, adding up to 735 sequences, of which 513 were of exactly 10 minutes, while 222 were less than 10 minutes.

The initially unavailable information on correctness of answers to quizzes became available later, leading to the addition of a new attribute, i.e. *score* that reflected the performance on all quizzes. Unlike the two attributes in the HTML-Tutor – number of correct and incorrect answers, the score attribute aggregates this information in one indicator (this is how it is logged in iHelp).

We also looked for two attributes to reflect the two types of disengagement behavior identified. As they seemed to be related to time, we intended to use the average time spent on each page across all users, as suggested by [22]. However, data analysis revealed that some pages are accessed by very small numbers of users, sometimes only one - a problem that was encountered in other research as well [23]. Consequently, we decided to use the average reading speed known to be in between 200 and 250 words per minute [24][25]. According to this reading speed, the majority of the pages would require less than 100 seconds (see Table 4) with only five pages exceeding 400 seconds.

TABLE 4
TIME INTERVALS FOR READING AND THE NUMBER OF PAGES IN EACH INTERVAL

Time interval	No of pages
500-550	3
400-500	2
300-400	5
200-300	41
100-200	145
<100	468

Some pages included images and videos that could increase the time needed to read/view the information displayed. However, only four of the 21 students attempted to watch videos and the number of attempts and their corresponding times per attempt and per student are displayed in Table 5.

Taking into account the above mentioned information about iHelp pages distribution, we defined a lower threshold of five seconds and an upper threshold of 420 seconds (seven minutes). The five seconds threshold for the minimal time to read a page seems to be a 'standard' in the literature (e.g. [23]). The 420 seconds threshold, even if somehow arbitrary, balances the factors involved in our particular case, namely:

1. Most pages, i.e. more than 99%, require less than 400 seconds to be read. Moreover, 70% of the

TABLE 5
NUMBER OF ATTEMPTS AND TIME SPENT WATCHING VIDEOS GROUPED BY USER

Subject	No of attempts	Time (sec.)
S1	1	3.47
S2	1	162
S3	9	1.16
	2	2.31
	1	94.91
S4	8	1.16
	2	2.31

pages require less than 100 seconds and only 5 pages, i.e. less than 1%, are left out.

2. Very few students watch videos (that could be longer than 5 or even 10 minutes, which would considerably affect the way to establish engagement level for a 10-minutes sequence)
3. There may be individual differences in reading speed and by allowing a rather loose upper threshold slow speed is taken into account. However, fast speed is not covered.
4. Some learners go through the material more than once, leading to an at least doubled time needed for reading.

Based on this analysis, the following two meta-attributes were defined: 1) NoPpP: the number of pages above the threshold established for maximum time required to read a page (420 seconds) and 2) NoPpM: the number of pages below the threshold established for minimum time to read a page (5 seconds). These two attributes were added for each sequence. We call them meta-attributes because they are derived from the raw data.

To account for the contribution of more instances and of the score attribute on one hand, and the contribution of the two new attributes (NoPpM and NoPpP) on the other hand, four dataset were defines. These are described in Table 6. By comparing datasets DS1_S2 and DS2_S2 (S2 stands for study 2) with datasets DS1_S1 and DS2_S1, the contribution of more instances and of the score attribute can be assessed; also this enables a more realistic comparison with the results from HTML-Tutor data. The results on datasets DS3_S2 and DS4_S2 will establish the influence of the two new attributes.

TABLE 6

DATASETS USED IN THE SECOND EXPERIMENT

Dataset	Sequences	Attributes
DS1_S2	All	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Score, Total time
DS2_S2	10 min.	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Score,Total time
DS3_S2	All	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Score, Total time, NoPpP, NoPpM
DS4_S2	10 min.	NoPages, AvgTimeP, NoQuestions, AvgTimeQ, Score,Total time, NoPpP, NoPpM

In the experiments, 68% of the sequences were used for training and 32% were used for testing. Also, the distribution of students within the two sets was controlled to avoid having sequences from the same user both in training and testing sets, which may introduce a positive bias in the results. The first study did not include this control, but the 10 iterations of the 10-fold cross-validation may have reduced the bias.

For the datasets including all 735 sequences (DS1_S2 and DS2_S2), 500 were used for training and 235 for testing. For the datasets with 10 minutes sequences only (DS3_S2 and DS4_S2), from the 513 instances, 348 were used for training and 165 for testing. The results are presented in Table 7.

TABLE 7

STUDY 2: EXPERIMENT RESULTS

Dataset	Results	BN	LR	SL	IBk	ASC	B	CvR	DT
DS1_S2	%correct	80.43	79.15	79.57	78.72	82.55	78.72	82.13	82.55
	TP rate	0.72	0.62	0.67	0.65	0.69	0.68	0.73	0.73
	FP rate	0.05	0.10	0.10	0.11	0.00	0.08	0.08	0.11
	Precision	0.80	0.85	0.82	0.81	0.88	0.79	0.83	0.84
	Error	0.28	0.27	0.36	0.27	0.28	0.29	0.27	0.27
DS2_S2	%correct	85.62	85.92	85.56	85.44	84.77	85.80	85.37	85.07
	TP rate	0.77	0.77	0.76	0.78	0.76	0.77	0.75	0.76
	FP rate	0.06	0.05	0.05	0.07	0.06	0.05	0.04	0.07
	Precision	0.93	0.93	0.94	0.91	0.92	0.93	0.94	0.92
	Error	0.23	0.21	0.22	0.20	0.24	0.23	0.24	0.23
DS3_S2	%correct	88.09	88.09	87.66	82.13	82.98	88.94	89.36	87.23
	TP rate	0.90	0.86	0.85	0.73	0.62	0.86	0.86	0.85
	FP rate	0.13	0.08	0.11	0.11	0.07	0.13	0.11	0.10
	Precision	0.92	0.87	0.87	0.85	1.00	0.89	0.90	0.86
	Error	0.16	0.20	0.22	0.19	0.25	0.15	0.15	0.17
DS4_S2	%correct	97.50	97.93	97.99	97.87	97.38	97.50	97.44	97.75
	TP rate	0.97	0.97	0.97	0.96	0.95	0.96	0.96	0.97
	FP rate	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.01
	Precision	0.98	0.99	0.99	0.99	1.00	0.99	0.99	0.99
	Error	0.04	0.03	0.04	0.03	0.05	0.04	0.04	0.03

Comparing the results from DS1_S2 and DS2_S2 with the results from study 1 (DS1_S1 and DS1_S2), an average increase of accuracy of 3.5% and 0.11%, respectively is noticed. The increase for the true positive rate is negligible, i.e. less than 0.3 for both datasets. Therefore, we can conclude that more data and the additional score attribute did not significantly improve the prediction results.

The results for DS1_S2 and DS2_S2 (the datasets without the new attributes) are lower compared to the results from the other two datasets (DS3_S2 and DS4_S2), indicating a positive influence of the two new attributes and a significant information gain. The percentage correct varies between 78% and 86%, while true positives rate has values between 0.62 and 0.78. Precision values range from 0.79 to 0.94; mean absolute error varies between 0.20 and 0.36.

The results for DS3_S2 and DS4_S1 (the datasets with the new attributes) presented in Table 7 show very good levels of prediction for all methods, with a correct prediction varying between approximately 82% and 98%. The results are similar for the true positives rate of the disengaged class, with most values varying between 0.85 and 0.97. However there are two deviant cases: for DS1_S2, the results obtained with IBk and ASC for the true positive rate are considerably lower, 0.73 and 0.62, respectively. Precision varies between 0.85 and 1.00 and error between 0.03 and 0.25.

As in the case of HTML-Tutor, the very similar results obtained from different methods and trials shows consistency of prediction and of the attributes used for prediction.

The highest percentage of correctly predicted instances was obtained using Simple Logistic classification on DS3_S2: 97.99%. The confusion matrix is presented in Table 8.

TABLE 8
THE CONFUSION MATRIX FOR SIMPLE LOGISTIC

		Predicted		
		Disengaged	**Engaged**	**Total**
	Disengaged	79	2	81
Actual	**Engaged**	0	84	84
	Total	79	86	165

Focusing on the disengaged learners we see that the same method, as well as other three methods (BN, LR, and DT), performs best on the same dataset: 0.97 true positives rate. The confusion matrix indicates that, on one hand, none of the engaged learners are classified as disengaged and, on the other hand, two disengaged learners are classified as engaged. Possible implications are that, in a real setting, engaged learners will not be interrupted for an intervention that is not required and that some disengaged learners will not be identified as such and, therefore, will not receive an intervention that would be required and beneficial.

Investigating the information gain of the attributes used in the analysis, the following ranking resulted from attribute ranking with information gain ranking filter as attribute evaluator: NoPpP, NoPages, AvgTimeP, NoPpM, AvgTimeQ, Score and NoQuestions.

The information gain brought by NoPpP is also reflected in the decision tree graph displayed in Fig. 1, where NoPpP is the attribute with the highest information gain, being the root of the tree. NoPpM also brings more information gain than attributes like Score and NoQuestions.

The ranking clearly indicates that attributes related to reading are more important that the ones related to taking quizzes. This is consistent with the structure of the learning environment that provides more material for reading than for testing. The two new attributes contribute with meta-information that improves the prediction results.

5 DISCUSSION

The two validation studies on iHelp data indicate that the attributes identified in the studies on HTML-Tutor data are relevant for the new system as well.

When comparing the results of iHelp studies with the HTML-Tutor data, lower true positive rates are noticed for the iHelp data. Given that the overall prediction rates are comparable, the difference may be accounted for by the different ways the two systems are used. While HTML-Tutor is freely accessible on the web, iHelp is used in a formal educational setting. This may account for the different percentage of disengaged instances in the two lots of data: 65% for HTML-Tutor and 49% for the iHelp.

The relatively low contribution of the *score* attribute came as a surprise, as intuitively, such information seems relevant for the prediction of engagement or disengagement. This is even more surprising when considering that such information is essential in related research focused on problem-solving activities. Nevertheless, this may indicate an important difference between problem solving environments and content delivering systems such as HTML-Tutor and iHelp where students engage in problem solving activities usually after having studied the related material. To look deeper into this issue, the ranking of attributes in HTML-Tutor and iHelp could be used to give us more information on the importance of such attributes in both systems. Before looking into this, we discuss the contribution of the two new attributes introduced in the second iHelp study: NoPpP (number of pages above the threshold of maximal reading time) and NoPpM (number of pages below the threshold of minimal reading time).

Comparing the DS4_S2 dataset from the second iHelp study (last from Table 7) containing the two new attributes with the HTML-Tutor results from Table 1, we notice an average increase of accuracy of 8.9% and an average increase of true positive rate of 0.03. This improvement is most likely accounted for by the two new attributes: NoPpP and NoPpM. The increase in the true positive rate

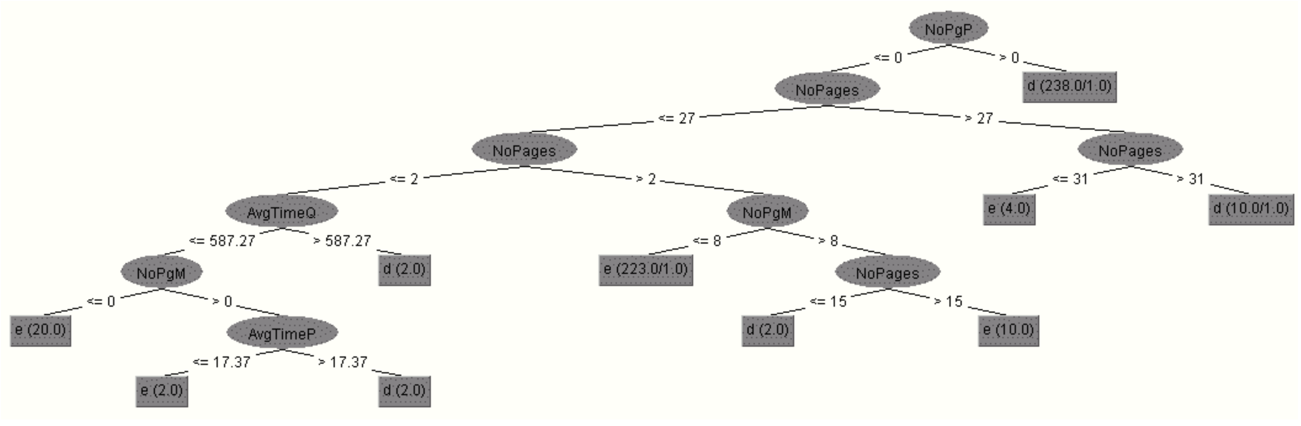


Fig. 1. Decision Tree for dataset DS4_S2

may not seem like a big improvement when directly compared with the HTML-Tutor results, but it is a considerable one when compared to the corresponding iHelp dataset without the two extra attributes, i.e. DS2_S2: 0.20 (or 20%). Therefore, the two new attributes significantly improve the prediction results.

To asses the contribution to prediction of the attributes in each system, three attribute evaluation methods with ranking as search method for attribute selection were used: chi-square, information gain and OneR [17]. For HTML-Tutor, according to chi-square and information gain ranking the most valuable attribute is *average time spent on pages*, followed by *number of pages*, *number of tests*, *average time spent on tests*, *number of correctly answered tests* and *number of incorrectly answered tests*. OneR ranking differs only in the position of the last two attributes: *number of incorrectly answered tests* comes before *number of correctly answered tests*.

The attribute ranking using information gain filter for iHelp attributes delivered the following ranking: NoPpP, NoPages, AvgTimeP, NoPpM, AvgTimeQ, Score, and NoQuestions. Chi-square evaluator produces the same ranking, except that the positions of the last two attributes are reversed, i.e. NoQuestions is before Score. OneR evaluator produces a different ranking compared to the other two, even if the main trend is preserved (attributes related to reading come before the one for quizzes): NoPpP, AvgTimeP, NoPages, NoPpM, NoQuestions, AvgTimeQ and Score. The comparison in Table 9 is based on information gain evaluator.

The attribute ranking results show that for both HTML-Tutor and iHelp, the attributes related to reading are more important than the ones related to tests. The iHelp *score* attribute and its two correspondent attributes from HTML-Tutor (*number of currently answered tests* and *number of incorrectly answered tests*) are among the least important ones.

Table 9 summarizes the similarities and dissimilarities between the findings from iHelp and HTML-Tutor studies. Although some differences exist, the main fact is that a good level of prediction obtained using similar attributes on datasets from two different systems and applying the same methods indicates that disengagement prediction is possible using information related to events like reading pages and taking tests (solving problems), information logged by most e-Learning system.

6 Perspectives and Conclusions

The validation studies suggest that our proposed approach for disengagement detection is potentially system independent and that it could be generalized to other systems. These results provide the blueprint for a component for automatic detection of disengagement that can be integrated into e-Learning systems to keep track of the learner's engagement status. Such a component offers the opportunity to intervene when appropriate – either automatically or through a tutor. We argue that disengagement detection represents the first step towards more detailed motivation elicitation. For example, once disengagement has been detected, the system may enter into a dialog with the learner in order to find out more about his/her motivation [26]. Furthermore, this information could be used for more targeted personalized intervention [27].

With both iHelp and HTML-Tutor two categories of disengaged learners were distinguished based on their patterns of behavior: a) disengaged students that click fast through pages without reading them and b) disengaged students that spend long time on a page, (far) exceeding the needed time for reading that page. Two of the previous approaches mentioned in Section 2 also present some patterns. Thus, we find a similarity between blind guess in [6] and unmotivated-guess in [9], on one hand, and the fast click through pages, on the other hand, as both reflect students' rush and lack of attention. However, we found no correspondent pattern in the literature for the long time spent on the same page. This may be due to the nature of the system, as this pattern is more likely to be displayed while reading rather than problem solving. This pattern also gives rise to problems like not knowing if a learner is disengaged with regards to the current activity and engaged in other behaviors like chatting with friends, reading email or using other software in general, or simply took an intentional break and spent the break time on the computer or somewhere else. This could easily be addressed by including in the system "break" and "resume" buttons for example. As the learners may forget to

TABLE 9

Similarities And Dissimilarities Between iHelp And Html-Tutor

	iHelp	HTML-Tutor
Prediction based on reading and tests attributes	85-86% with no additional attributes 97-98% with two additional attributes	85-91%
Attribute ranking (information gain)	1'. Number of pages above a threshold (NoPgP) 1. Number of pages accessed (NoPages) 2. Average time reading (AvgTimeP) 2'.Number of pages below a threshold (NoPgM) 3. Average time on quizzes (AvgTimeQ) 4. Score 5. Number of questions (NoQuestions)	1.Average time on pages 2. Number of pages 3. Number of tests 4. Average time on tests 5. Number of correctly answered tests 6. Number of incorrectly answered tests

use these buttons, another approach would be for the system to display a window after some time of inactivity asking the learner about the elapsed time was a break and if he would like some help. The help choice could trigger either a more detailed assessment of their motivation or an intervention strategy.

Despite the problem they may pose, knowledge about the two patterns of disengagement would be useful for a more targeted intervention and in further work the possibility to predict them will be investigated.

The two observed patterns of disengagement led to the introduction of two meta-attributes. Their usage considerably improved the prediction values. However, another way of using this knowledge would be to derive some rules that could be used for automatic annotations of data. For example, sequences for which the time spent on a page is above the upper threshold (420 seconds) for reading a page could be labeled as disengaged. Similarly, sequences that have more than two thirds of the pages below the lower threshold (5 seconds) for reading a page could be labeled as disengaged. This is another direction for future work that we intend to follow.

As already mentioned, previous research addressed disengagement and system gaming behavior [1], [10] (as a type of disengagement) only for problem-solving activities for which information on correctness or incorrectness of answers is very important, if not essential. For our approach this information has some importance, but it is not indispensable as shown in the first study on iHelp data. Therefore, if the learners are only reading, without doing any problem-solving activities, prediction of disengagement is still possible.

Moreover, the comparison of prediction values across the two validation studies on iHelp data suggests a limited impact on prediction of the amount of available data. The differences observed were quite small indicating the data necessary for training (at least for the initial one) is fairly modest, consequently facilitating the introduction of an automatic component for disengagement detection.

Acknowledgment

This work would not have been possible without access to the log data of the two learning systems NetCoach and iHelp. We would like to thank Gerhard Weber, University of Education Freiburg, Germany and Jim Greer, University of Saskatchewan, Canada, for their generous support.

References

[1] R. Baker, A. Corbett and K. Koedinger, "Detecting Student Misuse of Intelligent Tutoring Systems", *Proceedings of the Seventh International Conference on Intelligent Tutoring Systems*, pp. 531–540, 2004.

[2] T. Connolly and M. Stansfield, "Using Games-Based eLearning Technologies in Overcoming Difficulties in Teaching Information Systems", *Journal of Information Technology Education* 5, 459–476, 2006.

[3] G.D Chen, G.Y. Shen, K.L. Ou and B. Liu, "Promoting motivation and eliminating disorientation for web based courses by a multi-user game", *Proceedings of the EDMEDIA/ED-TELECOM 98 World Conference on Educational Multimedia and Hypermedia and World conference on Educational Telecommunications*, June 20-25, Germany (1998)

[4] C. R. Beal, L. Qu and H. Lee. "Classifying Learner Engagement through Integration of Multiple Data Sources". In *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 2–8, Menlo Park, California, 2006. AAAI Press.

[5] A. de Vicente and H. Pain, "Informing the Detection of the Students' Motivational State: an empirical Study", In S.A. Cerri, G., Gouarderes and F. Paraguau (eds.) *Intelligent Tutoring Systems, 6th International Conference*, pp. 933–943. Springer, Berlin, 2002.

[6] J. Beck, "Engagement tracing: Using response times to model student disengagement". In C. Looi, G. McCalla, B. Bredeweg and J. Breuker (eds.) *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pp. 88–95. IOS Press, Amsterdam, 2005.

[7] I. Arroyo and B.P. Woolf, "Inferring learning and attitudes from a Bayesian Network of log file data". In C.K. Looi, G. McCalla, B. Bredeweg and J. Breuker (eds.) *Artificial Intelligence in Education, Supporting Learning through Intelligent and Socially Informed Technology*, pp. 33–34. IOS Press, Amsterdam, 2005.

[8] L. Qu, N. Wang and W.L. Johnson, "Detecting the Learner's Motivational States in an Interactive Learning Environment". In C.-K. Looi et al. (eds.) *Artificial Intelligence in Education*, pp. 547–554. IOS Press, Amsterdam, 2005.

[9] J. Johns and B. Woolf, "A Dynamic Mixture Model to Detect Student Motivation and Proficiency". *Proceedings of the Twenty-first National Conference on Artificial Intelligence* (AAAI-06), Boston, MA, 2006.

[10] J. Walonoski and N.T. Heffernan, "Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems". In Ikeda, Ashley and Chan (eds.) *Proceedings of the Eighth International Conference in Intelligent Tutoring Systems*, pp. 382–391. Springer, Berlin , 2006

[11] J. Walonoski and N.T. Heffernan, "Prevention of Off-Task Gaming Behaviour within Intelligent Tutoring Systems". In Ikeda, Ashley and Chan (eds.) *Proceedings of the Eighth International Conference in Intelligent Tutoring Systems*, pp. 722–724. Springer-Verlag, Berlin, 2006.

[12] M. Cocea and S. Weibelzahl, "Eliciting Motivation Knowledge from Log Files towards Motivation Diagnosis for Adaptive Systems". In C. Conati, K. McCoy and G. Paliouras (eds.) *User Modelling 2007. Proceedings of 11th International Conference*, UM 2007, pp. 197-206, Springer, Berlin, 2007.

[13] P.R. Pintrich and D.H. Schunk, *Motivation in education: theory, research and applications.* Prentice Hall, Englewood Cliffs, 2002.

[14] J.M. Keller, "Development and use of the ARCS model of instructional design", *Journal of Instructional Development*, vol. 10, no. 3, pp. 2–10, 1987.

[15] G. Weber, H.-C. Kuhl and S. Weibelzahl, "Developing adaptive internet based courses with the authoring system NetCoach2. In *Hypermedia: Openness, Structural Awareness, and Adaptivity* (LNAI 2266), Springer, Berlin, pp. 226-238, 2001.

[16] M. Cocea and S. Weibelzahl, "Can Log Files Analysis Estimate Learners' Level of Motivation?", *Proceedings of ABIS Workshop, ABIS 2006 - 14th Workshop on Adaptivity and User Modeling in Interactive Systems*, Hildesheim pp. 32–35, 2006.

[17] I.H. Witten and E. Frank, *Data mining. Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kauffman Publishers, Elsevier, Amsterdam, 2005.

[18] T.M. Mitchell, *Machine Learning*. McGraw-Hill, New York, 1997.

[19] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol.20, No.1, pp.37–46, 1960.

[20] K. Krippendorff, *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: Sage, 2004.

[21] M. Lombard, J. Snyder-Duch and C. Campanella Bracken, "Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research" Retrieved on 06/11/06 from http://www.temple.edu/mmc/reliability. 2003.

[22] R. Rafter and B. Smyth, "Passive Profiling from Server Logs in an Online Recruitment Environment", *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization* (ITWP 2001) Seattle, Washington, USA, 2001.

[23] R. Farzan and P. Brusilovsky, "Social navigation support in E-Learning: What are real footprints", *Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization*, Edinburgh, U.K. pp. 49–56, 2005.

[24] ReadingSoft.com found at HYPERLINK,
http://www.readingsoft.com, 2007.

[25] TurboRead Speed Reading found at HYPERLINK,
http://www.turboread.com, 2007.

[26] M. Cocea, "Assessment of motivation in online learning environments". In V. Wade, H. Ashman and B. Smyth (eds.) *Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference*, AH 2006. LNCS, vol. 4018, pp. 414–418. Springer, Heidelberg, 2006.

[27] T. Hurley, "Intervention Strategies to Increase Self-efficacy and Self-regulation in Adaptive On-Line Learning". In V. Wade, H. Ashman and B. Smyth (eds.) *Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference*, AH 2006. LNCS, vol. 4018, pp. 440–444. Springer, Heidelberg, 2006.

Mihaela Cocea received her BSc in Psychology and Education from "Al. I. Cuza" University of Iasi in 2002 and her BSc in Computer Science from the same univerity in 2003. She has a taught MSc in Human Relations and Communication and completed her MSc by Research in Learning Technologies at National College of Ireland in 2007. She is currently working towards her PhD degree at the Department of Computer Science and Information Systems, Birkbeck College, University of London. Her research interests include intelligent learning environments, user modeling and adaptive feedback.

Dr Stephan Weibelzahl holds a lecturer position at the National College of Ireland in Dublin. He obtained his PhD from the University of Trier, Germany. After heading a research group at the Fraunhofer Institute of Experimental Software Engineering (IESE), Kaiserslautern, Germany, he joined National College of Ireland in 2004. With his background in psychology and computer science, he has long-standing research expertise in developing and evaluating Adaptive e-Learning Systems. His research interests include Adaptive Systems, learning technologies, evaluation, Knowledge Management and Blended Learning.