View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Portsmouth University Research Portal (Pure)

Log file analysis for disengagement detection in e-Learning environments

Mihaela Cocea, Stephan Weibelzahl

School of Informatics, National College of Ireland, Mayor Street, Dublin 1, Ireland
e-mail: mihaela@dcs.bbk.ac.uk, sweibelzahl@ncirl.ie

Abstract Most e-Learning systems store data about the learner's actions in log files, which give us detailed information about learner behaviour. Data mining and machine learning techniques can give meaning to these data and provide valuable information for learning improvement. One area that is of particular importance in the design of e-Learning systems is learner motivation as it is a key factor in the quality of learning and in the prevention of attrition. One aspect of motivation is engagement, a necessary condition for effective learning. Using data mining techniques for log file analysis, our research investigates the possibility of predicting users' level of engagement, with a focus on disengaged learners. As demonstrated previously across two different e-Learning systems, HTML-Tutor and iHelp, disengagement can be predicted by monitoring the learners' actions (e.g., reading pages and taking test/quizzes).
In this paper we present the findings of three studies that refine this prediction approach. Results from the first study show that two additional reading speed attributes can increase the accuracy of prediction. The second study suggests that distinguishing between two different patterns of disengagement (spending a long time on a page/test and browsing quickly through pages/tests) may improve prediction in some cases. The third study demonstrates the influence of exploratory behaviour on prediction, as most users at the first login familiarize themselves with the system before starting to learn.

Keywords e-learning, disengagement, log files analysis, educational data mining, motivation, user modelling

This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism

1. Introduction

Motivation is recognized as an important prerequisite of learning. While in a classroom setting motivation can be addressed by teachers, in e-Learning environments new ways of motivating or re-motivating learners are required. Several approaches addressing motivational issues have been proposed, including the design of attractive e-Learning systems (Ishii et al., 2004), using game features to motivate learners (Chen et al., 1998; Connolly & Stansfield, 2006), using whiteboards (Becta, 2002) and clickers (Martyn, 2007), as well as animated agents (Machado et al., 1999; Gussak & Baylor, 2003). Nevertheless, learners are not getting the full benefit of these features if they do not engage in the first place. These approaches focus on making the interaction attractive rather that addressing motivation in a personalised manner. Motivational issues often go beyond the facilities of a system and its engaging character to personal characteristics like the learners' attitudes to the subject matter, their attitudes toward the tutor (Beal et al., 2006) and their current mood (Beal & Lee, 2005). Therefore, knowledge about the engagement status and the motivational characteristics of learners could enhance the educational systems with detection capabilities and, ultimately, with personalised intervention strategies targeting the motivational status and characteristics of the learners.

Although there is no specific definition for engagement as a psychological concept, there are two theories that refer to it. One is Flow Theory (Csikszentmihalyi, 1997) and the other is the Theory of Engagement (Shneiderman et al., 1995). Flow Theory describes the state of flow which appears when several characteristics are met. Among these characteristics are: 1) clear goals; 2) concentrating and focusing; 3) balance between ability level and challenge; 4) a sense of personal control, etc. The second point, about concentration and focus, refers to engagement in the same meaning as used in our research.

The Theory of Engagement emerged in the mid-nineties in the context of teaching in electronic and distance education environments. The theory stresses the importance of being engaged in learning activities and the authors mention two ways of increasing engagement: collaboration and interaction with other learners, and meaningful tasks. The meaning of the term engagement is the same as the one previously mentioned, but the theory is focused on how to enhance it in the context of computer-supported learning environments.

To better understand the place of engagement in relation to motivation and other concepts associated with motivation (Pintrich & Schunk, 2002), we briefly describe the relation between engagement and some of these concepts: 1) engagement can be influenced by *interest*, as people tend to be more engaged in activities they are interested in; therefore, 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 summarized in the following way: 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 it is 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 one reason or another, the person is motivated to do the activity he/she is engaged in, or, conversely, if the person is disengaged, he/she may not be motivated to do the activity; in other words, engagement is an indicator of motivation.

We proposed a way of detecting disengagement in an unobtrusive way and we validated this approach across two e-Learning systems: HTML-Tutor and iHelp. During the

development of the approach we observed several aspects that could potentially improve it: 1) usage of reading speed attributes; 2) detection on two disengagement patterns: spending a long time on a page or test, and browsing fast through pages; and 3) eliminating exploratory sequences.

Our approach (Cocea & Weibelzahl, 2007a) was developed using data from HTML-Tutor. In the validation study (Cocea & Weibelzahl, 2007b) conducted on iHelp data, we introduced two attributes related to reading speed that considerably improved the prediction. These two reading speed attributes are related to a minimum and a maximum time required for reading a page. The effect of introducing these attributes on HTML-Tutor data is investigated in the first study presented in this paper; we expect an improvement of prediction performance.

The second study is very much related to the first, as the reading speed attributes somehow correspond to the two observed patterns of disengagement: long time spent on a page/test and fast browsing. The observation of these two patterns inspired the introduction of the reading speed attributes and thus, it is reasonable to assume that the usage of these attributes would improve the detection of the two patterns. Therefore, the purpose of the second study is twofold: i) to investigate the possibility of predicting the patterns of disengagement and ii) to identify the role of the corresponding attributes in the prediction, and, more specifically, to see if they improve prediction.

The third study aims to eliminate a possible bias of which we became aware by observing the log data. When learners first login to the system, they exhibit a somewhat "chaotic" and "illogical" behaviour as they are probably exploring the system before starting to use it. This exploratory behaviour is quite different from system usage behaviour and may have added a negative bias to our predictions. Hence, the third study investigates the influence on prediction of eliminating the exploratory sequences; we hypothesize an improvement of it.

The rest of the paper is organized as follows. Section 2 presents related research. Section 3 describes our approach to disengagement detection, including how we developed and validated it. Section 4 contains the three refinement studies briefly described above. The results are summarised and discussed in Section 5, and Section 6 concludes the paper.

2. Related Research

Several approaches for motivation detection from learner's interactions with e-Learning systems have been suggested, ranging from rule-based approaches and Item Response Theory models to Bayesian Networks. An overview of these approaches is presented in Table 1.

A rule-based approach based on ARCS Model (Keller, 1987) has been proposed to infer motivational states from the learners' behaviour using a ten-question quiz (de Vicente & Pain, 2002). 85 inference rules were produced by the participants who had access to replays of the learners' interactions with the system, as well as to the learner's motivational traits.

Another approach (Qu et al., 2005) based on the ARCS Model infers 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 plan (the learner's actions are compared to a learning plan) and number of questions asking for help.

A factorial analysis approach (Zhang et al., 2003) was used to group user's actions in two categories: actions that contribute to prediction of *attention* and actions that contribute to

prediction of *confidence*. The aspects targeted for prediction are two of the main concepts of the ARCS model.

Engagement tracing (Beck, 2005) 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 modelling disengagement; two methods of generating responses are assumed: blind guessing 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 (Johns & Woolf, 2006). 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 (called unmotivated-hint) and c) unmotivated and quickly guessing answers to find the correct answer (called unmotivated-guess).

Using a Bayesian Network, trained with log-data, variables related to learning and attitudes toward the tutor and the system can be inferred (Arroyo & Woolf, 2005). The log-data registered variables such as problem-solving time, mistakes and help requests.

A latent response model was proposed for identifying the students who try to game the system (Baker et al., 2004). Using a pre-test–post-test approach the gaming behaviour 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.

Table 1 Related research overview

Approach	Input	Output
Rule-based approach (de Vicente & Pain, 2002)	Learner's actions Motivational traits	Motivational states
Focus of attention (Qu et al., 2005)	Learner's focus of attention Current task Expected time to perform the task	Confidence Confusion Effort
Factorial analysis (Zhang et al., 2002)	Learner's actions	Attention Confidence
Engagement tracing (Beck, 2004)	Probability of correct response given a specific response time	Engagement Blind guessing
Dynamic mixture model (Johns & Woolf, 2006)	Student proficiency Evidence of motivation Student's response to a problem	Motivated Unmotivated-hint Unmotivated-guess
Bayesian network (Arroyo & Woolf, 2005)	Problem-solving time Mistakes Help requests	Attitudes to learning Attitudes towards the tutor Attitudes towards the system
Latent response model (Baker et al., 2004)	Student's actions Probabilities of prior skills	Harmful gaming Non-harmful gaming Non-gaming
Combined approach (Walonoski & Heffernan, 2006)	Classroom observations Logged actions	Guessing and checking Hint/ help abuse Non-gaming

The same problem of undesired gaming behaviour was addressed in (Walonoski & Heffernan, 2006a), an approach that combines classroom observations with logged actions in order to detect gaming behaviour manifested by guessing and checking or hint/help abuse.

Suggested prevention strategies (Walonoski & Heffernan, 2006b) include two different active interventions for the two types of gaming behaviour and a passive intervention. When the system detected that a student was exhibiting one of the two gaming behaviours, 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 of providing visual feedback on the student's actions and progress that was continuously displayed on the screen and available for viewing by the student and teacher.

All these approaches have the advantage of unobtrusively monitoring the learners' behaviour and identifying patterns associated with motivational issues. However they differ from our proposed approach in two aspects. First, the environments used include only problem-solving activities while we are interested in learning-type activities as well. Second, the domain is mathematics, which is rather technical and also a rather specialized domain which does not allow easy generalization to different areas; the domain considered in our research is HTML, which is at the border between technical and non-technical subjects and, therefore, may allow for an easier generalization across domains.

3. Motivation Modelling Framework

Our framework for motivation modelling includes two phases (Cocea, 2006). The first phase is related to disengagement and aims to identify the disengaged learners in an unobtrusive manner by monitoring their activity when using the system. The second phase includes a dialog with the learner that aims to get the learner involved in a self-assessment of several motivational characteristics related to Social Cognitive Learning Theory (Bandura, 1986). Using this two-step process, a complete motivational profile of the learner is obtained and personalized intervention can be delivered based on it. Moreover, monitoring the engagement status of the learner allows intervention at appropriate times.

In this paper we focus on the first phase of the modelling framework. More specifically, we explore three different ways of refining the disengagement detection approach. The steps taken in the development of this approach are briefly presented in the following subsections. They include: 1) a pilot study where we investigated the possibility of predicting disengagement from log files; 2) a main study where we identified the relevant actions for predicting disengagement; and 3) a validation study where we applied our approach to a second system in order to cross-validate the results.

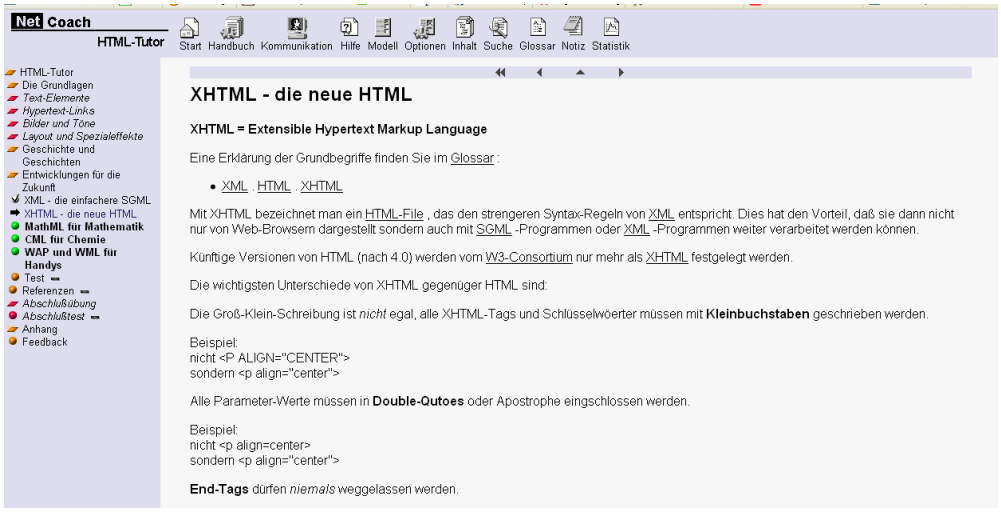
The first two studies used data from HTML-Tutor[1], a web interactive learning environment based on NetCoach (Weber et al., 2001). HTML-Tutor offers an introduction to HTML and publishing on the Web; it is online and can be accessed freely.

The tutorial is in German and contains seven chapters/high-level topics on HTML and publishing on the Web, e.g. text elements, hyperlinks, layout etc. The list of topics is displayed on the left of the screen – see Fig.1a. Each item in the list links to a file that is displayed in the main part of the screen. Several tools are accessible at the top of the screen among which are: a manual on how to use the system, communication tools, frequent questions, preferences on the way the information is displayed on the screen, a glossary of terms, taking notes and visualising statistics on the personal usage of the system (e.g. topics covered and performance on tests). A forward and back navigation bar is available under the tools bar and above the content. A guided tutorial about how to use the system is also available – see Fig.1b.

[1] http://art.ph-freiburg.de/HTML-Tutor/login-d.html

For the validation study, data from iHelp[2], the web-based learning environment from University of Saskatchewan, was used. iHelp includes two web-based applications designed to support both learners and instructors throughout the learning process: the iHelp Discussion system and iHelp Learning Content Management System or iHelp Courses. The iHelp Discussion system allows communication between students, between students and instructors and between students and subject matter experts.

(a)



(b)

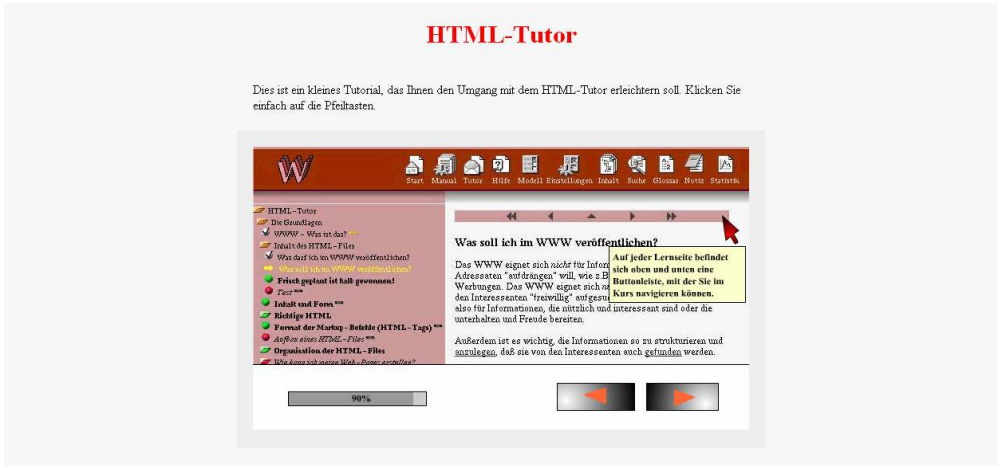


Fig.1 (a) HTML-Tutor screenshot; (b) Screenshot of the tutorial on how to use HTML-Tutor.

The iHelp Courses is designed for students working at a distance. It provides students with course content (text and multimedia), examples and quizzes/surveys. The content is organized in packages (containing hierarchical activities) with a single package displayed at a time on the left of the screen – see Fig.2. Each activity links to a file within the content package that is displayed in the main part of the screen. Forward and back navigation is available in the top right frame. The left hand menu included *course actions*, like preferences and search and *other actions* like logout. Access to collaboration tools, i.e. chat and

[2] http://ihelp.usask.ca/

discussion forum, is available at the bottom of the screen. The later is displayed in the bottom area of the screenshot in Fig.2.

Fig.2 iHelp Courses screenshot.

3.1. Pilot Study

A pilot study (Cocea & Weibelzahl, 2006) with a limited number of HTML-Tutor users has brought valuable information for the granularity of the timeframe for analysis. In this pilot study, we have used complete sessions as units of analysis; data from 20 learners (corresponding to 20 sessions) were analysed. Details on the methodology are given in the next section while here the focus is on findings that informed the design of the following study. Total time spent on the course turned out to be an important predictor of disengagement. However, according to this model a decision whether a learner is engaged or disengaged could only be made after 45 minutes, i.e. at a time when most disengaged students would have already left the system. We also noticed variation in the engagement level throughout a session: a learner could be engaged and then be disengaged for a while and engaged again and so on. We are interested in detecting disengagement and intervening appropriately before the learner leaves the system. We therefore decided to split learning session into sequences of 10 minutes for analysis.

The main purpose of this pilot study was to investigate the possibility of predicting disengagement from actions common to most e-Learning systems, like reading pages and taking tests. Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005) was used for analysis and decision trees method was chosen for its high interpretability. 75% correctly predicted instances (accuracy) for both engagement and disengagement, and 0.70 correctly identified instances (true positive rate) of disengagement, encouraged us to continue with our approach.

3.2. Prediction Model Development

The prediction model was developed on HTML-Tutor data. A list of all possible events that are recorded by HTML-Tutor is presented in Table 2. The second column displays the derived attributes used in the analysis for each of the possible events. For example, two

attributes related to reading pages are used: the number of pages and the average time spent reading pages. The other three columns display the range, mean and standard deviation for each attribute; all attributes referring to time are measured in seconds.

Table 2 Logged events and derived attributes used in the analysis with their range, mean and standard deviation

Events	Parameters/ Attributes	Range	Mean	Std. Dev.
Goal	The selected goal (from a list of 12 goals)	0-12	0.16	1.00
Preferences	Number	0-2	0.00	0.08
	Time spent selecting them	0-61	0.12	2.39
Reading pages	Number of pages	0-29	2.61	3.43
	Average time	0-600	270.73	258.49
Pre-tests	Number of pre-tests	0-34	0.21	2.19
	Average time	0-288	0.71	10.23
	Number of correct answers	0-33	0.17	1.92
	Number of incorrect answers	0-7	0.03	.371
Tests	Number of tests	0-31	2.01	3.78
	Average time	0-600	80.56	176.19
	Number of correct answers	0-28	1.34	2.65
	Number of incorrect answers	0-11	0.66	1.39
Hyperlink	Number of times accessed (frequency)	0-26	0.62	1.88
	Average time	0-600	31.64	98.94
Manual	Number of times accessed	0-2	0.01	0.10
	Average time	0-121	0.35	5.17
Help	Number of times accessed	0-2	0.01	0.13
	Average time	0-267	0.74	11.99
Glossary	Number of times accessed	0-5	0.10	0.39
	Average time	0-600	12.24	61.57
Communication	Number of times accessed	0-1	0.01	0.08
	Average time	0-2	0.00	0.08
Search	Number of times accessed	0-3	0.03	0.20
	Average time	0-600	13.84	89.19
Remarks	Number of times accessed	0-6	0.01	0.22
	Average time	0-113	0.14	3.73
Statistics	Number of times accessed	0-1	0.01	0.07
	Average time	0-159	0.53	8.25
Feedback	Number of times accessed	0-1	0.00	0.06
	Average time	0-18	0.05	0.83

Three human experts were involved in labelling the data with the level of engagement: one who classified all sequences (rater 1) and two (rater 2 and rater 3) who were involved in a coding reliability study presented below. They had access only to the unprocessed log files (split into sequences of 10 minutes) containing all events. In the pilot study, only two categories, engaged and disengaged, were used; however, due to the introduction of the sequences of 10 minutes of activity as the unit of analysis, the human raters occasionally had difficulty in deciding between the two, and thus a new category was introduced: neutral.

A small study was conducted to verify the reliability of the human coding. It included an *informal assessment* and an *additional expert rating*. The informal assessment was conducted using only 10 sequences; these were coded by rater 1 and rater 2. The ratings based on the given instructions were discussed to prevent different results due to instruction vagueness or suggestibility. The percent agreement between was 80% (only 2 different ratings from 10); the kappa measurement of agreement was 0.60 (p=.038) and the Krippendorff's alpha was

0.60 as well. In the *additional expert rating*, another expert (rater 3) coded 100 sequences randomly sampled from the 1015 entries in the dataset; the instructions used for the informal assessment were expanded with typical situations or patterns for each case. Details can be found in Cocea & Weibelzahl (2007a). The additional expert rating resulted in a rater agreement (between rater 1 and rater 3) of 92% (only eight different ratings from 100; in further discussion between the raters the eight disagreements were resolved) with a kappa measurement of agreement of 0.826 ($p<.01$) and Krippendorff's alpha of 0.8449. Although the percent agreement is high, we can see that kappa and Krippendorff's alpha have lower values. The percent agreement is not always the best indicator for agreement as it tends to be too liberal, while Cohen's Kappa and Krippendorff's alpha are known to be more conservative (Lombard et al., 2003). For the last two coefficients, values above 0.80 denote high inter-coder reliability, indicating that the engagement level within a 10 minutes sequence was established in an objective and reliable manner.

To establish whether there are significant differences in prediction levels due to certain attributes, three datasets (see Table 3) of HTML-Tutor data with different numbers of attributes were analyzed (Cocea & Weibelzahl, 2007a). The first dataset (DS-30) included all 30 attributes, the second dataset (DS-10) included 10 attributes related to reading pages, taking tests, following hyperlinks and consulting the glossary (these attributes were selected based on frequency of use by learners) and the third dataset (DS-6) included six attributes related only to reading pages and taking tests.

Table 3 HTML-Tutor datasets and attributes

Dataset	Attributes
DS-30	30 attributes related to all events
DS-10	10 attributes related to pages, tests (displayed below), hyperlinks (number of hyperlinks, average time) and glossary (number of times accessed, average time)
DS-6	6 attributes related to pages and tests (number of pages, average time on pages, number of tests, average time on tests, number of correctly answered tests, number of incorrectly answered tests)

The analysis included eight methods (Mitchell, 1997). These methods represent the most commonly used techniques for the data types of our datasets: nominal data for the predicted variable and numeric data for the predictors. All methods have the default Weka parameters unless specified otherwise. The eight methods described briefly are:

(a) Bayesian Nets with K2 algorithm and a maximum of 3 parent nodes (BN); Bayesian nets are popular in user modelling, having the advantage of providing a probability estimation rather than a threshold; also, they have shown high accuracy and speed when applied to large databases;
(b) Logistic regression (LR) models the probability of a categorical variable (e.g. disengagement in our case) occurring as a linear function of a set of predictor variables;
(c) Simple logistic classification (SL) uses the LogitBoost algorithm that performs additive logistic regression (combines several logistic regression models);
(d) Instance based classification with IBk algorithm (IBk) is a K-nearest neighbours classifier which is simple and effective; the nearest-neighbour method has been widely used in pattern-recognition since the 1960s;
(e) Attribute Selected Classification using J48 classifier and Best First search (ASC) combines the two methods referred to in its name: attribute selection and classification;

the dimensionality of training and test data is reduced by attribute selection before being passed on to the J48 classifier;

(f) Bagging using REP (reduced-error pruning) tree classifier (B); bagging is one of the methods used to improve classifier accuracy by combining results of several classifiers trials; the REP tree classifier is used because it is a fast decision tree learner (reduces the time required for the bagging to be performed);

(g) Classification via Regression (CvR) performs classification using regression methods; for each class of the predicted variable (engaged and disengaged in our case) a regression model is built;

(h) Decision Trees with J48 classifier based on Quinlan's C4.5 algorithm (Quinlan, 1993) (DT); the decision trees have the advantage of high interpretability and the possibility to convert to classification rules (e.g. to attach actions to certain situations corresponding to certain rules); however, even if they work well on relatively small datasets, scalability becomes an issue on large real-world databases.

Several stand-alone or combined prediction measurements are reported:

(a) percentage correct or accuracy:

$$Accuracy = \frac{\text{number of correctly classified instances}}{\text{total number of instances}}$$

The percentage of correct classifications shows how well the engagement level of the learners is accurately identified (for both engaged and disengaged).

(b) True Positive (TP) rate – illustrated for disengaged class:

$$TP\ rate = \frac{\text{number of correctly classified disengaged instances}}{\text{total number of disengaged instances}}$$

The True Positive rate for the disengaged class shows how well the disengaged learners are identified; it illustrates the correct classifications for the disengaged class.

(c) False Positive (FP) rate – illustrated for disengaged class:

$$FP\ rate = \frac{\text{number of incorrectly classified disengaged instances}}{\text{total number of disengaged instances}}$$

The False Positive rate for the disengaged class shows to what degree engaged learners are incorrectly predicted as disengaged; it illustrates the incorrect classifications for the disengaged class.

(d) Precision:

$$\text{Precision} = \frac{TP\ rate}{TP\ rate + FP\ rate}$$

Precision can be seen as a measurement of fidelity (closeness of repeated measures) for a given classification class (e.g. engaged or disengaged). High precision and low bias leads to high accuracy.

(e) Error:

$$Error = \frac{\text{number of incorrectly classified instances}}{\text{total number of instances}}$$

The error rate is the proportion of errors made over the whole of the test instances and, like accuracy, indicates the overall performance of a classifier.

The results are displayed in Table 4; they indicate a good level of prediction across all methods and datasets, with accuracy levels between 84% and 88%, and TP rate between 0.87

and 0.93. The mean absolute error varies between 0.10 and 0.15. The best overall prediction was obtained with the second dataset, i.e. DS-10: 88% correctly predicted instances, using Classification via Regression (CvR) and the best prediction for disengaged learners was 0.93 obtained using Bayesian Networks (BN).

The very similar results obtained from eight different data mining methods and using three different datasets indicates consistency of prediction and of the attributes used for prediction. The fact that there is small variation between the three datasets indicates that the most valuable attributes for predictions are the ones related to reading pages and taking tests (the only ones included in DS-6). These attributes also correspond to the most frequent actions within the HTML-Tutor. The fact that the best performance is obtained on DS-10 indicates that attributes related to hyperlinks and glossary contribute to a more accurate prediction. However, considering the Minimum Description Length principle (Witten & Frank, 2005), the frequency of events, the sparsity of data and the computational complexity, we argue for the use of the minimum number of attributes in the prediction model. Therefore, it should include only actions related to reading and tests: number of pages accessed, average time spent on pages, number of tests taken, average time spent on taking tests, number of correctly answered tests and number of incorrectly answered tests.

Table 4 Predictions of engagement level from HTML Tutor logs

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-30	**Accuracy**	87.07	86.52	87.33	85.62	87.24	87.41	87.64	86.58
	TP rate (disengaged)	0.93	0.93	0.93	0.92	0.93	0.93	0.92	0.93
	Precision (disengaged)	0.91	0.90	0.90	0.91	0.92	0.92	0.92	0.91
	Error	0.10	0.12	0.12	0.10	0.10	0.12	0.12	0.11
DS-10	**Accuracy**	87.18	85.88	85.82	85.13	86.03	86.87	88.07	85.16
	TP rate (disengaged)	0.93	0.93	0.93	0.91	0.92	0.92	0.91	0.91
	Precision (disengaged)	0.91	0.89	0.89	0.92	0.91	0.91	0.92	0.90
	Error	0.11	0.13	0.14	0.10	0.12	0.13	0.12	0.13
DS-6	**Accuracy**	86.68	84.15	84.05	85.18	86.95	86.90	87.21	86.20
	TP rate (disengaged)	0.93	0.93	0.93	0.90	0.92	0.92	0.91	0.92
	Precision (disengaged)	0.90	0.87	0.87	0.90	0.92	0.91	0.92	0.91
	Error	0.12	0.15	0.15	0.12	0.12	0.13	0.13	0.13

3.3. CROSS-SYSTEM VALIDATION

The next step was to cross-validate our prediction approach using a different e-Leaning system (Cocea & Weibelzahl, 2007b). Hence, we analyzed log files from an HTML course within iHelp, the web-based e-Learning system from University of Saskatchewan briefly described at the beginning of Section 3. We looked at 10 minute sequences, focussing on the same actions that were found most relevant in the previous experiment: reading pages and taking tests. Only 4 attributes were exactly the same as the ones used with HTML-Tutor: number of pages accessed, average time spent on pages, number of questions and average time spent on tests. Information on the correctness of answered questions was not available. Two new attributes related to reading speed were introduced: the number of pages exceeding the threshold established for maximum time required to read a page (420 seconds/7 minutes) and the number of pages below the threshold established for minimum time to read a page (5 seconds). The thresholds were established on the bases of the average reading speed and number of words per page; for details see Cocea & Weibelzahl (2007b).

Four datasets were used in the analysis: 1) DS-all+2 contains all attributes (including the two new ones – hence the notation '+2') and all sequences (including those of less than 10 minutes); 2) DS-all-2 was obtained by eliminating the two additional attributes from DS-all+2; 3) DS-600+2 contains all attributes, but only sequences of 10 minutes (600 seconds) (i.e. sequences of less than 10 minutes due to user log-off were skipped) and 4) DS-600-2 obtained by eliminating the two additional attributes from DS-600+2 (see Table 5). DS-all-2 and DS-600-2 were used to obtain a direct comparison with the HTML Tutor results. The other two datasets were used to investigate the impact of the two new attributes on prediction.

Table 5 iHelp datasets

Dataset	Sequences	Attributes
DS-all+2	All sequences	New attributes included
DS-all-2	All sequences	New attributes excluded
DS-600+2	10-minutes sequences only	New attributes included
DS-600-2	10-minutes sequences only	New attributes excluded

Table 6 Predictions of engagement level from iHelp logs

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-all+2	**Accuracy**	89.31	95.22	95.13	95.29	95.44	95.22	95.44	95.31
	TP rate (disengaged)	0.90	0.95	0.95	0.94	0.94	0.94	0.95	0.95
	Precision	0.90	0.95	0.95	0.96	0.97	0.97	0.96	0.96
	Error	0.13	0.07	0.10	0.05	0.08	0.08	0.08	0.07
DS-all-2	**Accuracy**	81.73	83.82	83.58	84.00	84.38	85.11	85.33	84.38
	TP rate (disengaged)	0.78	0.82	0.81	0.79	0.77	0.79	0.80	0.78
	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
DS-600+2	**Accuracy**	94.65	98.06	97.91	98.59	97.65	97.65	97.76	97.47
	TP rate (disengaged)	0.95	0.97	0.96	0.98	0.96	0.96	0.96	0.96
	Precision	0.94	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Error	0.07	0.02	0.04	0.02	0.05	0.04	0.03	0.03
DS-600-2	**Accuracy**	84.29	85.82	85.47	84.91	84.97	85.38	85.26	85.24
	TP rate (disengaged)	0.78	0.77	0.76	0.77	0.75	0.76	0.75	0.75
	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

The same system, i.e. Weka, and the same eight data mining methods were used. The results are displayed in Table 6. For the DS-all-2 and DS-600-2 similar results to HTML-Tutor are observed: accuracy rates are between 82% and 85% while TP rate varies from 0.75 to 0.82. If, for the accuracy, there is a slight decrease, for the TP rate the decrease is higher. This may be due to the lack of attributes related to quizzes/surveys results.

For the DS-all+2 and DS-600+2 the results vary between 89% and 98% for the accuracy and between 0.90 and 0.99 for the TP rate, indicating that the two new attributes improve the prediction level. The best overall prediction was 98.56%, obtained using Instance Based Classification with IBk algorithm on DS-600+2. The best disengagement prediction was 0.98 using the same method and the same algorithm. This is not surprising since this method is known to be simple and effective (Witten & Frank, 2005).

The similarity of results for HTML-Tutor and iHelp obtained using similar attributes and the same methods indicates that engagement prediction is possible using information related to reading pages and taking tests, information logged by most e-Learning system. Hence, we

can conclude that our proposed approach for engagement prediction is system independent and could be generalized to other systems.

However, while designing the approach we identified aspects that could improve the prediction even more. The pursued path was 'construction' (using HTML-Tutor data), followed by validation (using iHelp data). As mentioned in the introduction, we observed two patterns of behaviour with HTML-Tutor which proved to be present with iHelp as well. This led to the introduction of two new attributes related to reading speed and to the idea of investigating the prediction of the two patterns. Thus, we introduced the attributes in the validation study and observed a considerable improvement of prediction. Hence, the following study followed the reversed path, i.e. from the validation study back to the 'construction' in order to verify whether the two additional attributes improve prediction in HTML-Tutor data as well. Very much related to this is the idea of verifying if the separation of different types of engagement improves prediction. This investigation is supported not only by our observations on the data, but also from literature, several studies reporting usage of different categories of 'motivational status'. Another observation on data from both HTML-Tutor and iHelp led to the idea of the third study: at the first login on the system, learners adopted an exploratory behaviour which is different from the following behaviour that can be characterized as usage of the system. Thus, we also investigate whether the elimination of the exploratory behaviour improves prediction.

4. Disengagement Prediction Refinement

This section includes the three studies introduced previously: 1) reading speed attributes validation study; 2) patterns of disengagement prediction study; and 3) elimination of exploratory sequences study.

4.1. Validation Of Reading Speed Attributes

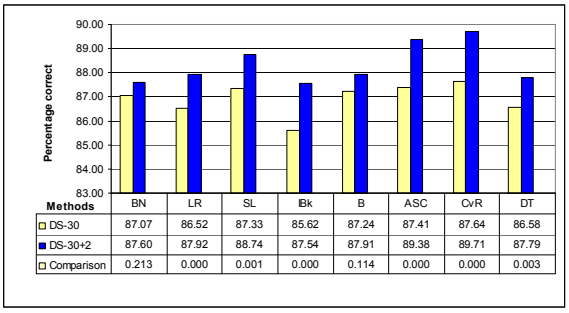
For each sequence of 10 minutes in the HTML-Tutor log data, the two attributes used with iHelp were added: the number of pages exceeding the 420 second threshold and the number of pages below five seconds. We compared the predictions obtained after adding these attributes with the predictions obtained without them. All three datasets, DS-30, DS-10 and DS-6, were included. The study design is presented in Table 7. Our hypothesis is that the two additional attributes will improve the overall and especially the disengagement prediction level.

Table 7 Validation of reading speed attributes study design

	30 attributes	**10 attributes**	**6 attributes**
With original attributes	DS-30	DS-10	DS-6
With the 2 additional attributes	DS-30+2	DS-10+2	DS-6+2

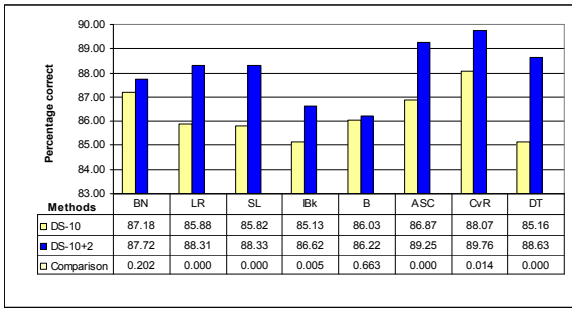
As in the 'construction' study, log files of 48 subjects were used; they spent between 1 and 7 sessions on HTML-Tutor, each session varying between 1 and 92 sequences. The dataset included 1015 entries (i.e. sequences), of which 943 were of exactly 10 minutes and 72 varied between 7 and 592 seconds. The datasets used in this study included only the 943 entries of exactly 10 minutes.

As in the previous studies, Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005) was used for analysis and the eight methods presented in Section 3.2 were considered; the experiment utilized 10-fold stratified cross-validation iterated 10 times. We were interested in the quality of the predictions in terms of two classification parameters: the percentage of correct classifications (accuracy) as well as the true positive rate for disengaged. The results are grouped in six figures: the first three (Fig. 3, Fig. 4 and Fig. 5) display the comparison for the accuracy, while the next three (Fig. 6, Fig. 7 and Fig. 8) display the comparison for the true positive rate for disengaged.



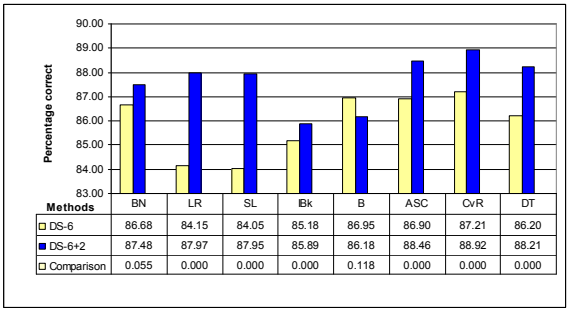
Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-30	87.07	86.52	87.33	85.62	87.24	87.41	87.64	86.58
DS-30+2	87.60	87.92	88.74	87.54	87.91	89.38	89.71	87.79
Comparison	0.213	0.000	0.001	0.000	0.114	0.000	0.000	0.003

Fig. 3 Accuracy for original dataset with 30 attributes (DS-30) and the same dataset with the two additional attributes (DS-30+2)



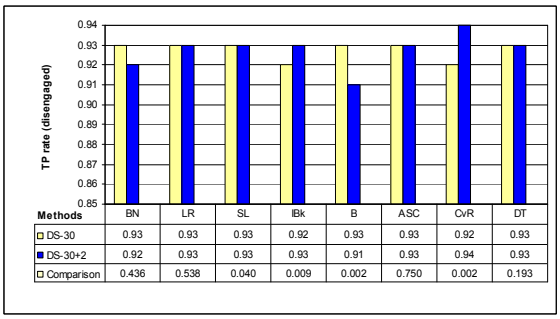
Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-10	87.18	85.88	85.82	85.13	86.03	86.87	88.07	85.16
DS-10+2	87.72	88.31	88.33	86.62	86.22	89.25	89.76	88.63
Comparison	0.202	0.000	0.000	0.005	0.663	0.000	0.014	0.000

Fig. 4 Accuracy for original dataset with 10 attributes (DS-10) and the same dataset with the two additional attributes (DS-10+2)



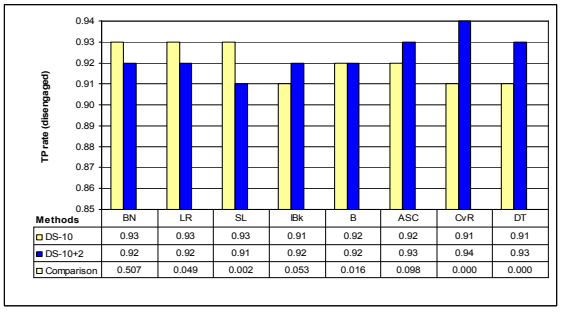
Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-6	86.68	84.15	84.05	85.18	86.95	86.90	87.21	86.20
DS-6+2	87.48	87.97	87.95	85.89	86.18	88.46	88.92	88.21
Comparison	0.055	0.000	0.000	0.000	0.118	0.000	0.000	0.000

Fig. 5 Accuracy for original dataset with 6 attributes (DS-6) and the same dataset with the two additional attributes (DS-6+2)



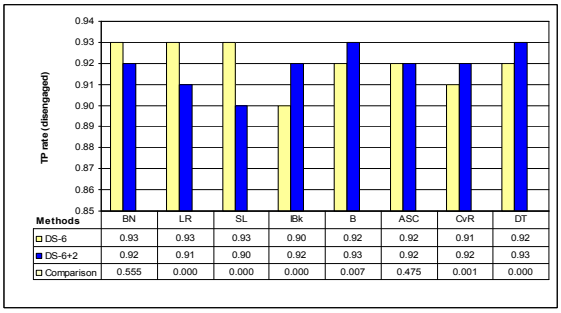
Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-30	0.93	0.93	0.93	0.92	0.93	0.93	0.92	0.93
DS-30+2	0.92	0.93	0.93	0.93	0.91	0.93	0.94	0.93
Comparison	0.436	0.538	0.040	0.009	0.002	0.750	0.002	0.193

Fig. 6 TP rate (disengaged) for original dataset with 30 attributes (DS-30) and the same dataset with the two additional attributes (DS-30+2)



Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-10	0.93	0.93	0.93	0.91	0.92	0.92	0.91	0.91
DS-10+2	0.92	0.92	0.91	0.92	0.92	0.93	0.94	0.93
Comparison	0.507	0.049	0.002	0.053	0.016	0.098	0.000	0.000

Fig. 7 TP rate (disengaged) for original dataset with 10 attributes (DS-10) and the same dataset with the two additional attributes (DS-10+2)



Methods	BN	LR	SL	IBk	B	ASC	CvR	DT
DS-6	0.93	0.93	0.93	0.90	0.92	0.92	0.91	0.92
DS-6+2	0.92	0.91	0.90	0.92	0.93	0.92	0.92	0.93
Comparison	0.555	0.000	0.000	0.000	0.007	0.475	0.001	0.000

Fig. 8 TP rate (disengaged) for original dataset with 16 attributes (DS-6) and the same dataset with the two additional attributes (DS-6+2)

For the first three database pairs, there are significant differences for 6 out of 8 methods: LR, SL, IBk, ASC, CvR and DT. In all cases, the accuracy is higher for the databases with

the two additional attributes. Therefore, we consider that in the case of overall prediction, our hypothesis was confirmed. The accuracy increase is statistically significant. To demonstrate that we applied either the paired t-test if both result sets were normally distributed, or the Wilcoxon test otherwise. The normal distribution was verified using the Kolmogorov-Smirnov test. All differences were significant ($p<.014$) with only two exceptions. First, the observed decrease in Bagging using REP (B) was not significant ($p>.05$) for all three data sets; the fact that bagging predictors are rather stable against perturbations of the data sets (Breiman 1996) might explain this lack of improvement. Second, the increase for Bayesian Nets (BN) also did not reach a statistically significant level; this may be due to the influence of the ordering of the attributes on the K2 algorithm (the reading attributes were added last). To further investigate this aspect, the algorithm could be run with different (random) orderings.

For applying these algorithms in a diagnosis of real learners, the true positive rate is of importance as well: while identifying disengaged learners is critical for enabling appropriate intervention, not classifying engaged learners as disengaged by mistake is important for not interrupting engaged learners. As shown in Fig. 6, Fig. 7 and Fig. 8, the impact of the additional attributes varies across data sets and classification methods. While there is no significant change for BN and ASC, CvR (for all datasets) and DT (for DS-10 and DS-6) significantly improve. The remaining methods show an inconsistent picture with both increases and decreases. All changes are relatively minor, with a maximum improvement of .03 and a maximum reduction of .02.

In two situations, Fig. 6: SL and Fig. 7: B, it appears in the graph that the true positive rate for the two databases (DS-30 and DS-30+2 in Fig. 6; DS-10 and DS-10+2 in Fig. 7) has the same value: 0.93 in Fig. 6 and 0.92 in Fig. 7. At the same time for these cases it appears that the differences for each of the two pairs of databases are significant. This is explained by the fact that the figures displayed are rounded to two digits. The four digit values are: for Fig. 6, DS-30: 0.9343; Fig. 6, DS-30+2: 0.9267, Fig. 7, DS-10: 0.9153 and Fig. 7, DS-10+2: 0.9248.

The results suggest a trade-off between accuracy and the true positive rate, especially for logistic regression and simple logistic classification, for which we observed a significant increase in accuracy and a significant decrease of the true positive rate. This involves, on the one hand, a better detection of engagement (hence, the increase in accuracy) and, on the other hand, an increase of misclassification of disengaged instances as engaged (false positive rate) (hence a decrease of true positive rate).

In summary, the two new reading speed attributes improve the accuracy of classification, but have mixed effects for the true positive rate. While the differences in terms of accuracy are stable across different methods and statistically significant, the nominal improvement is limited (.01–.04). Correlation via Regression shows the biggest improvement in the true positive rate (relevant for diagnosis) and the best overall performance with up to 89.8% correct classifications and a true positive rate of up to 0.94. Considering that the two new attributes are supposed to be indicators of disengagement, the results are somehow surprising: we expected a stronger effect on the true positive rate for the disengaged class when using the new attributes. Trying to clarify these results, in the study conducted for prediction of the two patterns, we also looked at the impact of the attributes on the predictions.

4.2. Disengagement Patterns Prediction

Disengagement, in fact, comprises at least two different types of behaviour. The experts who rated the sequences reported that some learners seem to spend a very long time on a single

page, while others seem to click through pages without reading. We designed a study to investigate whether predicting these two patterns of disengagement as opposed to a single disengagement state would improve the prediction model. We labelled these patterns as follows: 1) fast browsing through pages/tests was denoted as "disengaged-fast" (DF) and 2) long time spent on the same page/test was denoted as "disengaged-long" (DL). Although these names are not expressing opposite situations, their names were chosen because they express the corresponding behaviour most accurately. The investigation was conducted with both HTML-Tutor and iHelp.

In the light of the results of the reading speed study, we decided to run two trials: with and without the two additional reading speed attributes. This would enable us to cross-compare the results of the two studies.

4.2.1. *Disengagement Patterns in HTML-Tutor*

The study design was very similar to the previous study. We started from the same six datasets (used in the validation of speed attributes study), but used four levels of engagement: engaged, neutral, "disengaged-long", and "disengaged-fast". To distinguish the datasets used in this study compared to the previous one, the "L/F" label was added on the names of the datasets to indicate that "disengaged-long" and "disengaged-fast" patterns are used.

The sequences were coded as "disengaged-long" or "disengaged-fast" using the same rules as the ones used in the validation study briefly presented in Section 3.3: if in a sequence the learner spent more that 420 seconds (7 minutes) on a page or test, the sequence was coded "disengaged-long" (DL); if in a sequence 2/3 of the total number of pages were below 5 seconds, the sequence was coded "disengaged-fast" (DF).

The same maximum threshold was used as with iHelp because all pages from HTML-Tutor require less than 400 seconds to be read. The minimum threshold, 5 seconds, was also the same; this threshold has been used in other studies (e.g. Farzan & Brusilovsky, 2005), and there seems to be an agreement about this minimal time to process the information on a page, regardless if the time is spent to read the page or to look for other links.

From the total of 943 sequences of 10 minutes, 646 were DL and only 21 DF. Thus, as there were too few instances of DF, we focused on the DL pattern. The same software and methods were used for the analysis; 10-fold cross-validation iterated 10 times was applied. Table 8 shows the accuracy and the TP rate for DL for all datasets. In order to see whether there are significant differences between the two distributions, we applied the same procedure as in the validation of reading speed attributes study.

Good accuracy levels were obtained, with values between 85.2% and 89.2%, which are slightly lower than the ones obtained when disengagement was only one category (see Table 4) and also slightly lower than the ones presented in the validation of reading speed attributes study (Fig. 3, Fig. 4 and Fig. 5, Section 4.1). This was expected due to the introduction of the two patterns. On the other hand, the TP rates for "disengaged-long", with the two additional attributes, with values between 0.89 and 0.95, are higher than the previous results from both the prediction model development study and the reading speed attributes validation study (Fig. 6, Fig. 7 and Fig. 8, Section 4.1).

To have a picture of the predictions across all trials, some distributions of accuracy and TP rates are displayed in Fig. 9, Fig. 10 and Fig. 11. These distributions give an idea of the most frequent accuracy levels and help identify situations in which a relatively good level of prediction is obtained from compact and consistent predictions rather than an average between poor and very good prediction levels.

Table 8 HTML Tutor predictions of engagement levels when the two disengaged patterns, DL and DF, are considered; true positive rate is displayed only for DL.

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-30-L/F	**Accuracy**	84.33	86.31	87.14	84.66	87.12	86.81	87.16	86.10
	TP rate	0.89	0.94	0.95	0.93	0.95	0.94	0.94	0.94
DS-30+2-L/F	**Accuracy**	86.68	87.50	88.32	85.82	87.68	88.27	89.12	87.53
	TP rate	0.93	0.95	0.94	0.94	0.93	0.94	0.95	0.95
DS-10-L/F	**Accuracy**	83.40	85.96	85.69	84.37	86.66	86.37	87.47	85.20
	TP rate	0.88	0.93	0.94	0.92	0.93	0.94	0.94	0.92
DS-10+2-L/F	**Accuracy**	86.94	87.63	87.96	85.80	85.83	88.65	89.22	88.27
	TP rate	0.94	0.93	0.93	0.93	0.95	0.95	0.95	0.95
DS-6-L/F	**Accuracy**	83.06	83.90	84.00	82.41	86.95	86.52	86.73	85.86
	TP rate	0.89	0.92	0.93	0.91	0.93	0.93	0.93	0.92
DS-6+2-L/F	**Accuracy**	86.33	87.01	87.16	85.16	85.97	87.81	88.44	87.83
	TP rate	0.94	0.92	0.91	0.94	0.95	0.94	0.94	0.95

Fig. 9 displays the distribution of the accuracy for the best method (CvR) on DS-6-L/F. Most values are between 86% and 93%, and all of them are above 81%. What appears like vertical axes in the graph is a result of the fact that values are based on fractional percent of the 95 test cases; for example 85/95 is approximately 89%. More common results for a certain value of accuracy are visible in the higher frequency of dots along the vertical lines (e.g. most frequent values are around 89%).

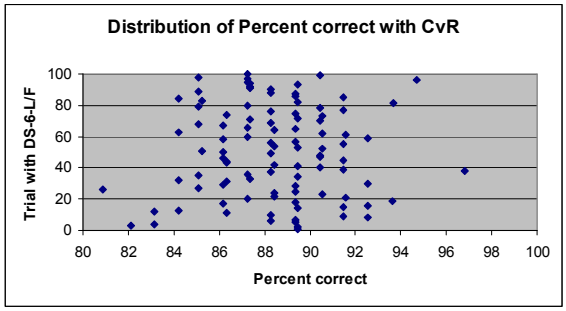


Fig. 9 Distribution of accuracy with CvR on DS-6-L/F**.**

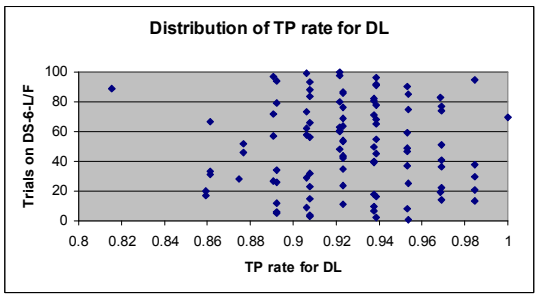


Fig. 10 Distribution of TP rate for DL using CvR on DS-6-L/F (without the two additional attributes).

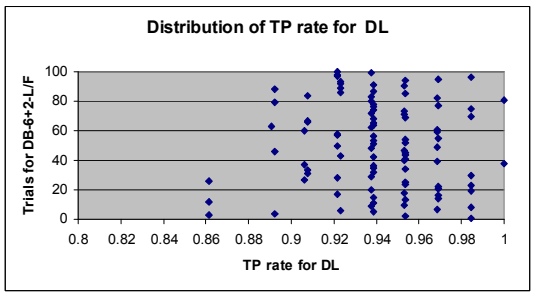
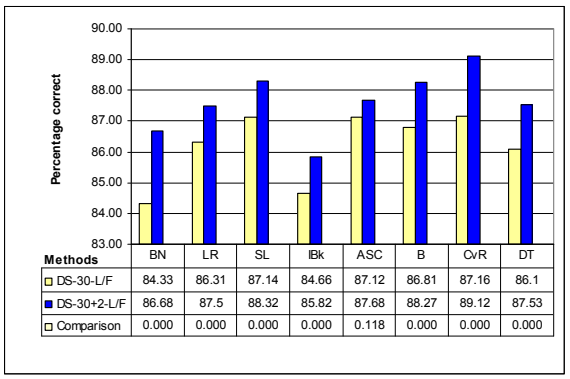


Fig. 11 Distribution of TP rate for DL using CvR on DS-6+2-L/F (with the two additional attributes).

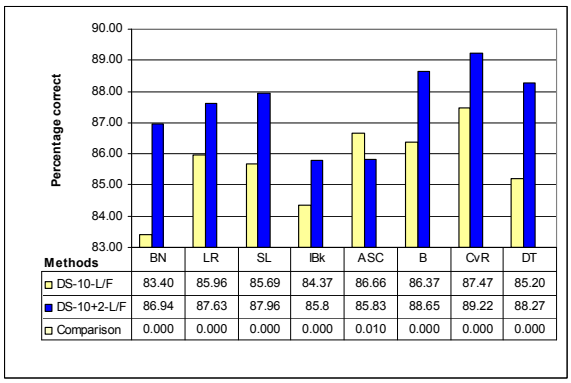
For the TP rate for "disengaged-long" on DS-6-L/F and DS-6+2-L/F using CvR we notice close values: 0.94 with the two additional attributes and 0.93 without them. The graphs displayed in Fig. 10 and Fig. 11 show that the distributions have more or less the same range,

with the exception of an outlier around the value of 0.81 in Fig. 10. However, the values are distributed differently, with a higher density of larger numbers when the two attributes are used.



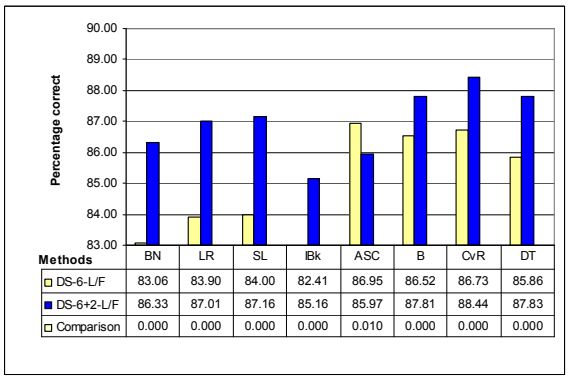
Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-30-L/F	84.33	86.31	87.14	84.66	87.12	86.81	87.16	86.1
DS-30+2-L/F	86.68	87.5	88.32	85.82	87.68	88.27	89.12	87.53
Comparison	0.000	0.000	0.000	0.000	0.118	0.000	0.000	0.000

Fig. 12 Accuracy comparison between DS-30-L/F and DS-30+2-L/F.



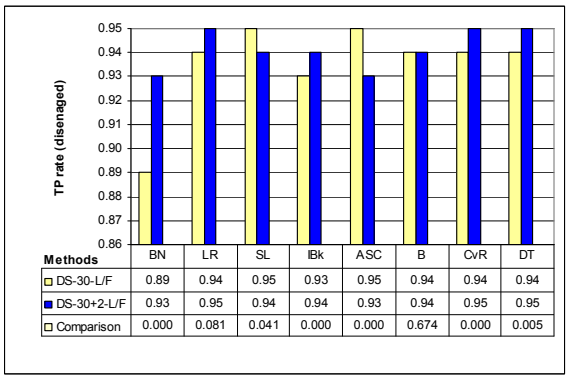
Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-10-L/F	83.40	85.96	85.69	84.37	86.66	86.37	87.47	85.20
DS-10+2-L/F	86.94	87.63	87.96	85.8	85.83	88.65	89.22	88.27
Comparison	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000

Fig. 13 Accuracy comparison between DS-10-L/F and DS-10+2-L/F



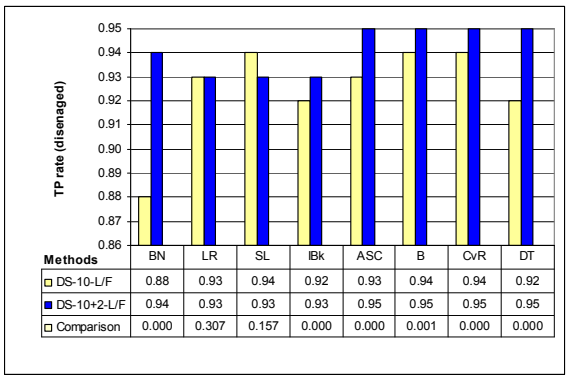
Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-6-L/F	83.06	83.90	84.00	82.41	86.95	86.52	86.73	85.86
DS-6+2-L/F	86.33	87.01	87.16	85.16	85.97	87.81	88.44	87.83
Comparison	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000

Fig. 14 Accuracy comparison between DS-6-L/F and DS-6+2-L/F.



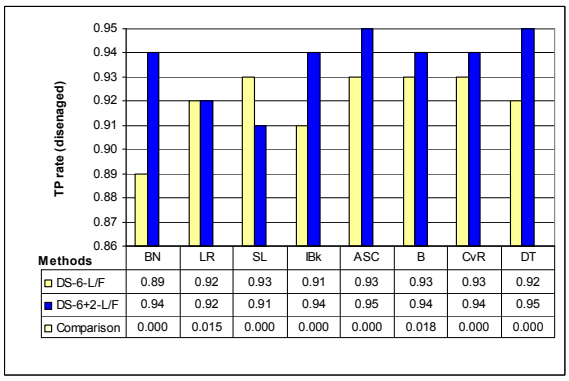
Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-30-L/F	0.89	0.94	0.95	0.93	0.95	0.94	0.94	0.94
DS-30+2-L/F	0.93	0.95	0.94	0.94	0.93	0.94	0.95	0.95
Comparison	0.000	0.081	0.041	0.000	0.000	0.674	0.000	0.005

Fig. 15 TP rate for DL comparison between DS-30-L/F and DS-30+2-L/F.



Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-10-L/F	0.88	0.93	0.94	0.92	0.93	0.94	0.94	0.92
DS-10+2-L/F	0.94	0.93	0.93	0.93	0.95	0.95	0.95	0.95
Comparison	0.000	0.307	0.157	0.000	0.000	0.001	0.000	0.000

Fig. 16 True positive rate for DL comparison between DS-10-L/F and DS-10+2-L/F.



Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-6-L/F	0.89	0.92	0.93	0.91	0.93	0.93	0.93	0.92
DS-6+2-L/F	0.94	0.92	0.91	0.94	0.95	0.94	0.94	0.95
Comparison	0.000	0.015	0.000	0.000	0.000	0.018	0.000	0.000

Fig. 17 True positive rate for DL comparison between DS-6-L/F and DS-6+2-L/F.

The datasets with the two additional reading speed attributes show higher accuracy than the corresponding datasets without the attributes – see Fig. 12, Fig. 13 and Fig. 14. All differences are statistically significant.

Most TP results are higher for these datasets as well. In a few cases, in particular for DS-6, the differences did not reach statistical significance. All SL results, one LR (DS-6) and one ASC (DS-30) result show a decrease in TP rate.

For LR in Fig. 17 it appears that the values are the same, although the difference is significant; as in a previous situation this is explained by the fact that the results were rounded to two digits. Looking at the values with four digits, for DS-6-L/F the value is 0.9238 and for DS-6+2-L/F is 0.9150. Consequently, for LR the true positive rate is higher for the dataset without the new attributes.

In summary, separating the two disengagement patterns results, on the one hand, in a decrease in accuracy of up two 5%. On the other hand, the true positive rate for DL increased for most methods. (An equivalent analysis for DF was not possible due to the small number of cases.) In other words, in an on-line course where it is important to identify people who spend too much time on single pages, it may be worthwhile to separate out "disengaged-long" behaviour from "disengaged-fast" behaviour. However, this comes with the cost of incorrectly classifying some engaged learners as disengaged. Overall, it would be recommended to use the reading speed attributes in connection with the DF/DL classification, as it improves accuracy as well as TP rate in most cases. This is not surprising as DF/DL is certainly related to reading speed.

4.2.2. *Disengagement Patterns in iHelp*

In order to cross-validate the results in a second e-Learning system, we replicated the study using the iHelp data. However, only the datasets with the new reading speed attributes were used: DS-all+2 and DS-600+2. As we don't need to distinguish them from the ones without the additional attributes, '+2' was eliminated from the notation. In order to distinguish the datasets used in this study from the following one, we added "L/F" to indicate that "disengaged-long" and "disengaged-fast" patterns are included.

From the total of 450 sequences, 169 were DL and 82 were DF. DS-all-L/F includes all instances, while DS-600-L/F includes only sequences of exactly 10 minutes (340 with 161 DL and 8 DF). Both datasets include all attributes. Since DS-600-L/F contained only 8 DF instances, we investigated only the overall and DL prediction on this dataset. The larger number of DF instances in DS-all-L/F compared to DS-600-L/F indicates that the learners that are "disengaged fast" tend to spend less than 10 minutes on the system; they also tend to occur before the learner leaves the system and, consequently, this pattern may indicate that a learner is about to logout.

The same tool and methods were used, as well as the 10-fold stratified cross-validation iterated 10 times. The results are presented in Table 9.

An additional measurement is presented – d prime – that indicates how well the disengagement levels can be distinguished. D-prime is a measure used especially in signal theory to judge how well signals can be distinguished from noise. The d-prime formula adapted to statistical notation is:

$$d' = z(TP) - z(FP)$$

The z-transform function has the role of transforming measures with different ranges of absolute values to a common scale to allow comparison. This function has a normal distribution with the mean value set to 0 and the range of most values is within 3 standard deviations above and below the mean.

D-prime values above 2 show that engagement levels can be accurately distinguished and identified.

Table 9 iHelp predictions of engagement levels with the two disengaged patterns, DL and DF.

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-all-L/F	**Accuracy**	89.27	91.13	91.13	88.87	88.98	90.22	90.62	89.73
	TP rate DL	0.91	0.92	0.91	0.92	0.91	0.91	0.91	0.91
	FP rate DL	0.01	0.02	0.02	0.04	0.02	0.01	0.02	0.02
	d'	3.67	3.46	3.39	3.16	3.46	3.67	3.46	3.46
	TP rate DF	0.73	0.84	0.85	0.76	0.74	0.79	0.81	0.80
	FP rate DF	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	d'	2.36	2.75	2.79	2.46	2.39	2.56	2.63	2.59
	Error	0.11	0.09	0.09	0.08	0.11	0.10	0.10	0.10
DS-600-L/F	**Accuracy**	93.14	94.58	94.40	94.13	93.76	93.90	94.28	93.81
	TP rate DL	0.93	0.94	0.93	0.94	0.93	0.93	0.93	0.93
	FP rate DL	0.02	0.03	0.02	0.04	0.02	0.02	0.02	0.03
	d'	3.53	3.44	3.53	3.31	3.53	3.53	3.53	3.53
	Error	0.08	0.06	0.07	0.05	0.08	0.07	0.07	0.07

Similar to the HTML-Tutor study, accuracy results are pretty high for both datasets, ranging between 88.9% and 94.6%. The TP rate for DL is also high with values from 0.91 to 0.94; the FP rate for DL ranges from 0.01 to 0.04. The TP rate for DF has unexpectedly high values between 0.73 and 0.84, while the FP rate for DF goes from 0.02 to 0.04; the error ranges from 0.05 to 0.11. In all cases the smaller dataset with sequences of exactly 10 minutes (DS-600-L/F) exceeded the complete dataset (DS-all-L/F). The d prime values are extremely good for both DL and DF, indicating a good discrimination of both patterns. The distribution of accuracy on DS-all-L/F for one of the bests performing methods, SL, is presented in Fig. 18, where we can see that most values fall between 86% and 96%. These values are lower than the original results (see Table 4) where no distinction between the two disengagement patterns was made. The distribution of TP rate for DL includes values from 0.70 to 1 (Fig. 19), with most values above 0.86. Again, compared to the original results, the prediction performance decreased.

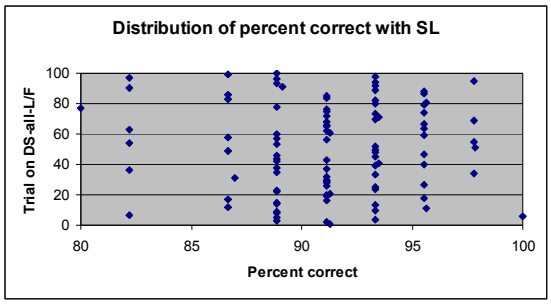


Fig. 18 Distribution of accuracy with SL on DS-all-L/F.

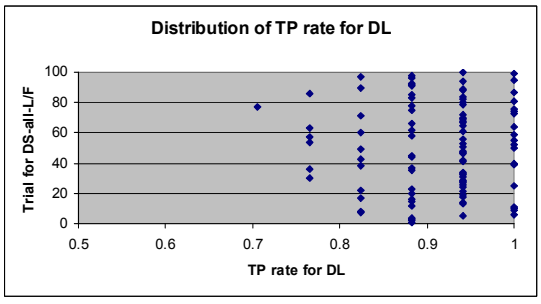


Fig. 19 Distribution of TP rate for DL on DS-all-L/F, using SL method.

Fig. 20 displays the distribution of TP rates for DF. The results obtained were unexpectedly high, with most values above 0.75, with the highest density of values around

0.88 and with 19 cases (out of 100) with value 1, meaning exact prediction. Considering the low number of instances for DF, these values were surprising and encouraging.

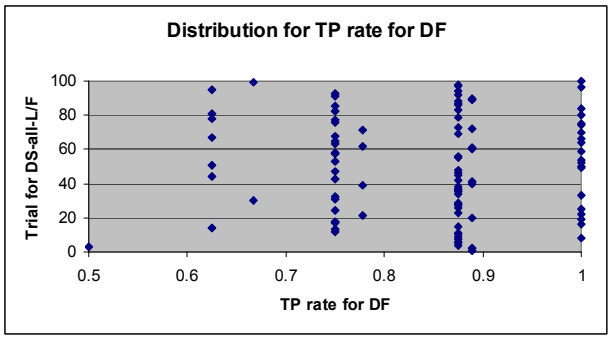


Fig. 20 Distribution of true positives rate for DF on DS-all-L/F, using SL method.

In summary, the iHelp data confirms the results of the HTML-Tutor study. The introduction of the two disengagement patterns led to a small decrease, i.e. around 3%, for the overall prediction. However, the prediction values are still very good; moreover, a good discrimination has been shown for both patterns, "disengaged-long" and "disengaged-short", suggesting that the disengagement patterns should be used in on-line courses where the identification of the two types of disengaged learners is of particular importance.

4.3. Exclusion Of Exploratory Patterns

Besides the two patterns investigated in the previous study, we observed with both HTML-Tutor and iHelp that on the first login to the system, the learners tend to behave in an exploratory manner; they click on the menu options and on the links to the main chapters of the course in a rather "chaotic" way. This familiarising behaviour is different from what was observed with the following sequences, when the learners seem to focus on the content. Given this difference, the presence in the analysis of the initial sequences where the exploratory behaviour occurs may negatively influence the results. This study was conducted in order to explore the influence of the exclusion of these exploratory sequences on prediction values. Both systems, HTML-Tutor and iHelp, were considered.

4.3.1. *HTML-Tutor*

From the 943 sequences, the 65 of them representing the first sequence of the first session, were eliminated. Consequently, the dataset used for analyses included 878 instances. We included all datasets, i.e. DS-30, DS-10 and DS-6, with (labelled "dl/df/e/n") or without (labelled "d/e/n") the two patterns; all datasets contained the reading speed attributes.

The results are displayed in Table 10. The accuracy values vary between 86% and 89%, while the TP rates for DL range from 0.91 to 0.96; the FP rates are from 0.14 to 0.23 and the error values are between 0.07 and 0.11. The d prime values are above 2 for all datasets indicating a good discrimination of the "disengaged-long" pattern in all of them.

Table 10 HTML-Tutor: Prediction results without the exploratory sequences

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-30+2 (d/e/n)	**Accuracy**	87.82	89.07	89.53	87.47	87.82	89.21	89.53	88.21
	TP rate d	0.93	0.94	0.94	0.93	0.94	0.93	0.94	0.93
	FP rate d	0.21	0.23	0.20	0.20	0.23	0.18	0.18	0.21
	Error	0.10	0.10	0.10	0.09	0.11	0.10	0.10	0.10
	d'	2.28	2.29	2.40	2.32	2.29	2.39	2.47	2.28
DS-30+2 (dl/df/e/n)	**Accuracy**	86.83	88.39	88.83	85.73	87.59	88.84	89.35	88.92
	TP rate DL	0.93	0.95	0.95	0.93	0.95	0.95	0.95	0.96
	FP rate DL	0.19	0.20	0.18	0.18	0.21	0.16	0.16	0.16
	Error	0.08	0.08	0.08	0.07	0.09	0.09	0.08	0.08
	d'	2.35	2.49	2.56	2.39	2.45	2.64	2.64	2.75
DS-10+2 (d/e/n)	**Accuracy**	87.81	88.43	88.31	87.35	88.06	89.18	89.45	88.19
	TP rate d	0.93	0.93	0.93	0.92	0.94	0.93	0.94	0.93
	FP rate d	0.21	0.21	0.19	0.19	0.22	0.18	0.18	0.21
	Error	0.10	0.10	0.11	0.09	0.11	0.11	0.10	0.10
	d'	2.28	2.28	2.35	2.28	2.33	2.39	2.47	2.28
DS-10+2 (dl/df/e/n)	**Accuracy**	86.83	87.98	88.01	85.98	87.59	88.92	89.34	88.88
	TP rate DL	0.93	0.94	0.94	0.93	0.95	0.95	0.96	0.96
	FP rate DL	0.19	0.21	0.17	0.16	0.21	0.16	0.16	0.16
	Error	0.08	0.08	0.08	0.07	0.09	0.09	0.08	0.08
	d'	2.35	2.36	2.51	2.47	2.45	2.64	2.75	2.75
DS-6+2 (d/e/n)	**Accuracy**	87.76	87.89	87.53	86.10	88.01	88.23	88.52	87.86
	TP rate d	0.93	0.92	0.91	0.93	0.94	0.93	0.92	0.94
	FP rate d	0.23	0.18	0.17	0.23	0.22	0.19	0.17	0.24
	Error	0.11	0.11	0.11	0.10	0.11	0.11	0.11	0.11
	d'	2.21	2.32	2.30	2.21	2.33	2.35	2.36	2.26
DS-6+2s (dl/df/e/n)	**Accuracy**	87.06	87.35	87.44	84.59	87.63	88.19	88.64	88.35
	TP rate d	0.94	0.93	0.93	0.93	0.95	0.94	0.94	0.96
	FP rate DL	0.21	0.16	0.14	0.20	0.21	0.17	0.16	0.18
	Error	0.09	0.09	0.09	0.08	0.09	0.09	0.09	0.09
	d'	2.36	2.47	2.56	2.32	2.45	2.51	2.55	2.67

Looking at accuracy, we observe the following:

1) compared to results from the validation of reading speed attributes (no patterns included): a) Datasets with all attributes: the values are more or less the same, with four cases where the values are higher when the exploratory sequences are considered, and four cases where the values are higher with exploratory sequences excluded (see Fig. 21); b) Datasets with 10 attributes: the values are lower when the exploratory sequences are included for half of the cases, i.e. four out of eight (see Fig. 22); c) Datasets with 6 attributes: the values are higher when the exploratory sequences are included for five out of eight cases (see Fig. 23);
2) compared to the results from the pattern detection study: a) Datasets with all attributes: the values are higher when the exploratory sequences are excluded, in six cases out of eight (see Fig. 24); b) Datasets with 10 attributes: the values are higher when the exploratory sequences are excluded for most cases - seven out of eight (see Fig. 25); c) Datasets with 6 attributes: the same situation as for the datasets with 10 attributes (see Fig. 26).

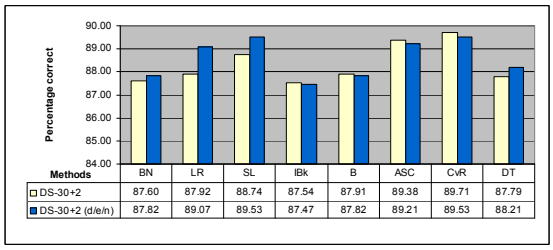


Fig. 21 Accuracy comparison between DS-30+2 and DS-30+2 (d/e/n).

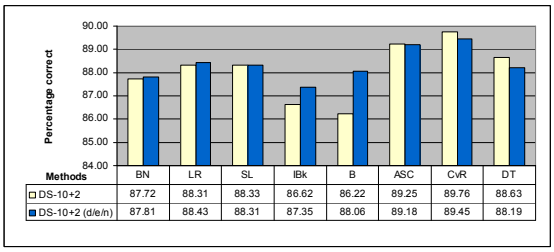


Fig. 22 Accuracy comparison between DS-10+2 and DS-10+2 (d/e/n).

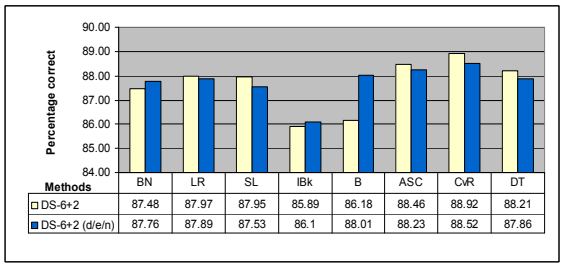


Fig. 23 Accuracy comparison between DS-6+2 and DS-6+2 (d/e/n).

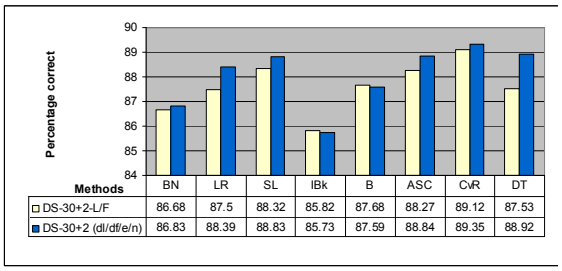


Fig. 24 Accuracy comparison between DS-30+2-L/F and DS-30+2 (dl/df/e/n).

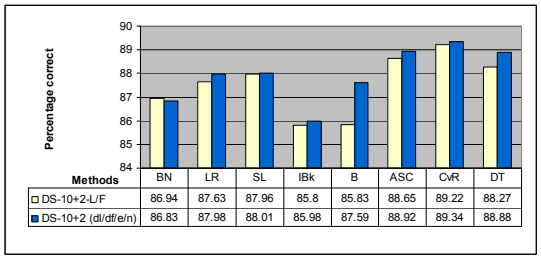


Fig. 25 Accuracy comparison between DS-10+2-L/F and DS-10+2 (dl/df/e/n).

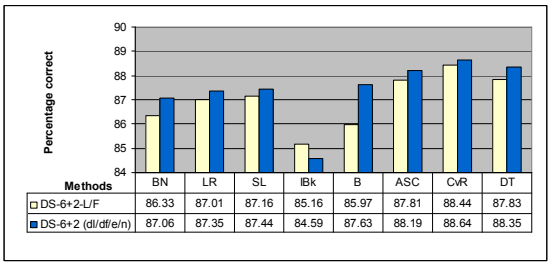


Fig. 26 Accuracy comparison between DS-6+2-L/F and DS-30+2 (dl/df/e/n).

Looking at the true positive rate for disengaged and respectively, "disengaged-long", we observe the following:

1) compared to results from the validation of the reading speed attributes (TP for disengaged): a) Datasets with all attributes: the values are the same in four cases and in the other four the values are higher when the exploratory sequences are excluded (see Fig. 27); b) Datasets with 10 attributes: the same situation as for the datasets with 30 attributes (see Fig. 28); c) Datasets with 6 attributes: in one case the values are the same; for the other seven the values are higher when the exploratory sequences are excluded (see Fig. 29);
2) compared to the results from the patterns detection study (TP for "disengaged-long"): a) Datasets with all attributes: the values are higher when the exploratory sequences are excluded in four cases; in one case the opposite situation is encountered; for the other three cases, the values are the same (see Fig. 30); b) Datasets with 10 attributes: the same situation as for the datasets with 30 attributes (see Fig. 31); c) Datasets with 6 attributes: in four cases the values are the same; in one case the value is higher when the exploratory sequences are included; for the remaining three cases, the values are higher when the exploratory sequences are excluded (see Fig. 32).

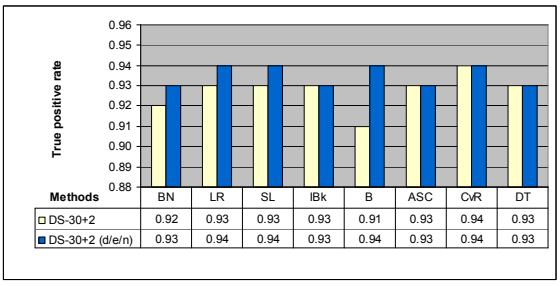


Fig. 27 TP rate for disengagement (d) comparison between DS-30+2 and DS-30+2 (d/e/n).

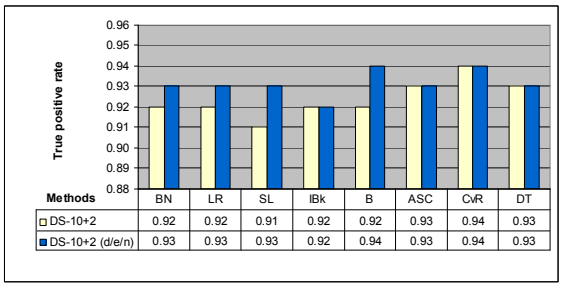


Fig. 28 TP rate for disengagement (d) comparison between DS-10+2 and DS-10+2 (d/e/n).

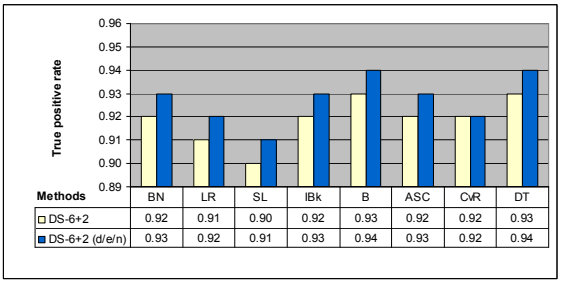


Fig. 29 TP rate comparison for disengagement (d) between DS-6+2 and DS-6+2 (d/e/n).

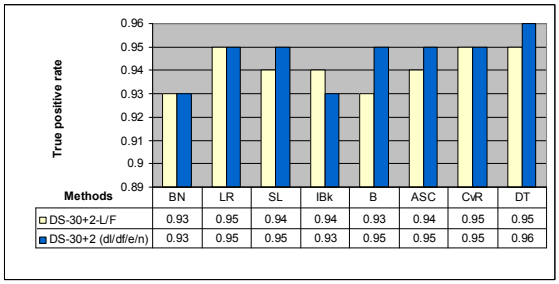


Fig. 30 TP rate for DL comparison between DS-30+2-L/F and DS-30+2 (dl/df/e/n).

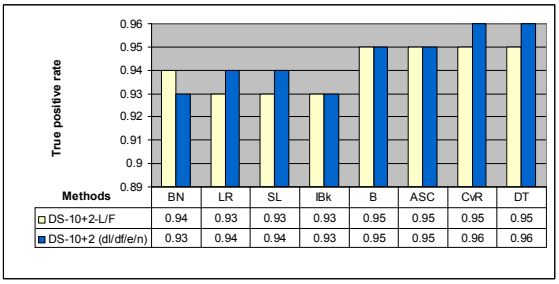


Fig. 31 TP for DL comparison between DS-10+2-L/F and DS-10+2 (dl/df/e/n).

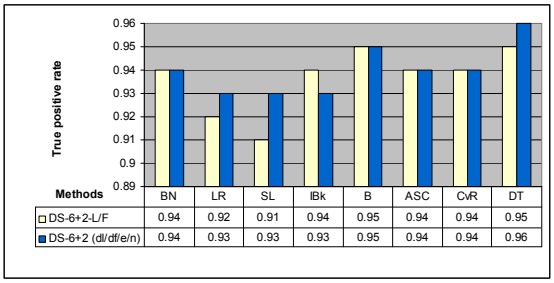


Fig. 32 TP rate for DL comparison between Ds-6+2-L/F and DS-6+2 (dl/df/e/n).

Summarizing the results, if the accuracy was more-or-less the same with or without the exploratory sequences, the TP rate for disengaged and "disengaged-long" improved in most cases when the exploratory sequences were excluded. This indicates that excluding the exploratory sequences positively influences the prediction and suggests the training should not include the exploratory sequences.

4.3.2. *iHelp*

Like in the previous study with iHelp data, two datasets were used: DS-all including all sequences and DS-600 including only sequences of exactly 10 minutes. Both datasets included the reading speed attributes and the two patterns of disengagement: "disengaged-long" and "disengaged-fast". To distinguish these datasets from the ones used in the patterns of disengagement study, "dl/df/e" was added to the names of the datasets.

From dataset DS-all, 11 exploratory sequences were excluded, while from DS-600 only 3 such sequences were eliminated. This indicates that in 8 cases out of 11 the learners spent less than 10 minutes on their first login to the system. The results are displayed in Table 11. They show good levels of accuracy, between 88% and 95%, TP rates for DL between 0.92

and .94 and TP rates for DF (only for DS-all) from 0.70 and 0.81. The FP rates for DL vary between 0.01 and 0.04, while the ones for DF range from 0.04 to 0.05; the error has values between 0.05 and 0.08. The d prime values over 2 indicated a good discrimination of both "disengaged-long" and "disengaged-fast" patterns.

Comparing the results presented in Table 11 with the ones from the patterns of disengagement study (Table 9, Section 4.2.2) and focusing on accuracy, we observe the following: 1) for DS-all: in five cases the values are higher when the exploratory sequences are included and in three cases the values are higher when the exploratory sequences are excluded – see Fig. 33; 2) for DS-600: for all eight methods the values are higher when the exploratory sequences are excluded – see Fig. 34.

Table 11 iHelp: Prediction results without the exploratory sequences

		BN	LR	SL	IBk	ASC	B	CvR	DT
DS-all (dl/df/e)	**Accuracy**	88.48	91.48	91.46	88.79	89.18	90.16	90.53	89.57
	TP rate DL	0.92	0.93	0.92	0.92	0.92	0.92	0.92	0.92
	FP rate DL	0.01	0.02	0.02	0.04	0.02	0.02	0.02	0.02
	d'	3.73	3.53	3.46	3.16	3.46	3.46	3.46	3.46
	TP for DF	0.71	0.81	0.81	0.70	0.71	0.75	0.76	0.76
	FP for DF	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	d'	2.20	2.63	2.63	2.28	2.30	2.43	2.46	2.46
	Error	0.08	0.07	0.07	0.06	0.09	0.08	0.07	0.08
DS-600 (dl/df/e)	**Accuracy**	93.53	94.98	94.63	94.47	94.06	94.27	94.66	94.33
	TP rate DL	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	FP rate DL	0.02	0.02	0.02	0.03	0.02	0.02	002	0.02
	Error	0.07	0.06	0.07	0.05	0.07	0.07	0.06	0.06
	d'	3.53	3.61	3.61	3.44	3.61	3.61	3.61	3.61

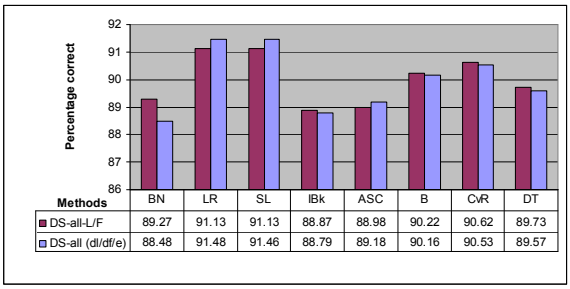


Fig. 33 Accuracy comparison between DS-all-L/F and DS-all (dl/df/e).

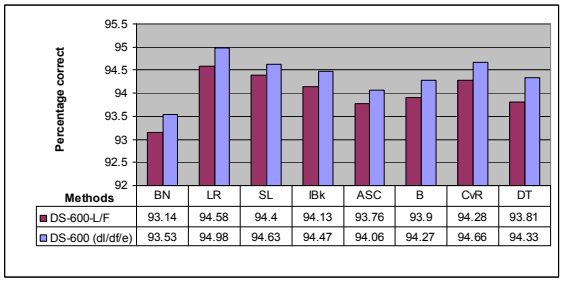


Fig. 34 Accuracy comparison between DS-600-L/F and DS-600 (dl/df/e).

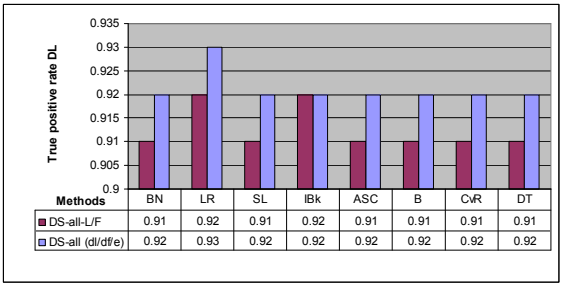


Fig. 35 TP rate for DL comparison between DS-all-L/F and DS-all (dl/df/e).

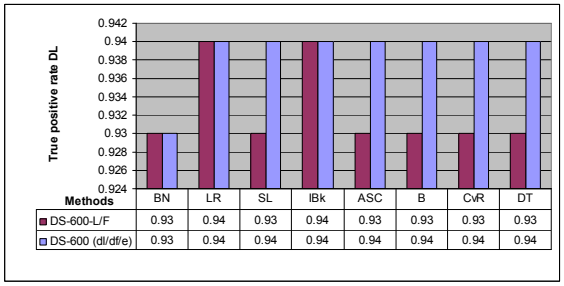
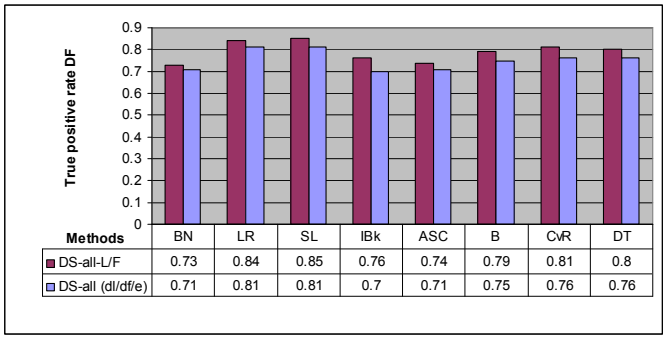


Fig. 36 TP rate for DL comparison between DS-600-L/F and DS-600 (dl/df/e).

Comparing the results from Table 11 with the ones from the patterns of disengagement study (Table 9, Section 4.2.2) and focusing on true positive rate for DL, the following can be observed: 1) for DL-all: in seven cases the values are higher when the exploratory sequences are excluded, and in one case the values are the same – see Fig. 35; 2) for DS-600: for five methods the values are higher when the exploratory sequences are excluded and for the other three the values are the same – see Fig. 36.

Comparing the results from Table 11 with the ones from the patterns of disengagement study (Table 9, Section 4.2.2) and focusing on true positive rate for DF (only dataset DS-all), a decrease is observed for all methods when the exploratory sequences are excluded – see Fig. 37.



Methods	BN	LR	SL	IBk	ASC	B	CvR	DT
DS-all-L/F	0.73	0.84	0.85	0.76	0.74	0.79	0.81	0.8
DS-all (dl/df/e)	0.71	0.81	0.81	0.7	0.71	0.75	0.76	0.76

Fig. 37 TP rate for DF comparison between DS-all-L/F and DS-all (dl/df/e).

Summarizing the results from this study, on one hand, we observe an increase for accuracy and true positive rate for DL and, on the other, a decrease for the true positive rate for DF when the exploratory sequences are excluded. Considering that from the 11 exploratory sequences that were eliminated, 10 were DF, the fact that the already small number of DF sequences was reduced even more may explain the decrease in the true positive rate for DF. However, the elimination of these sequences brought an increase of the overall predictive values and of the ones for DL, suggesting that the exclusion of the exploratory sequences may be more beneficial when detection of the DL pattern is of particular interest.

The two studies presented previously deal only with exploratory behaviour occurring at the very beginning of the interaction with the system, when the exploratory behaviour most frequently occurred – in our case the first sequence of the first session. However, for some users this behaviour could occur for less than 10 minutes, for more than 10 minutes or at subsequent logins to the system, i.e. at the beginning of each session or even during sessions. Therefore, the elimination of the exploratory behaviour only from the beginning does not solve the whole problem. Moreover, from an educational point of view, the exploratory behaviour, although different from the general usage of the system, is not necessarily a deviation from learning. This aspect is discussed in more detail in the next section.

5. Summary and Discussion

The studies presented in this paper describe and refine a general approach for disengagement prediction for e-Learning systems. We argue that disengagement detection will play a vital role in the development of personalized e-Learning environments that adapt to motivational

characteristics of learners. The most frequent characteristic used for adaptation in e-Learning, so far, is knowledge; other learner characteristics often used include learning goals, interests, preferences, and others. However, these characteristics are influenced by learner's motivation, and considering them separately, leads to an incomplete and potentially inaccurate view of the learner.

As a step towards this aim we developed an approach for prediction of disengagement as one aspect of motivation that will help identify unmotivated learners that often have learning difficulties. This, in turn, will facilitate personalised interventions that would take into account their motivational status as well as other characteristics.

In this paper we presented three refinement studies aiming to improve our approach to disengagement detection. An overview of the key results is presented in Table 12.

Table 12 Overview of the advantages and disadvantages of the different refinement strategies for disengagement detection

Refinement	Description	Advantages	Disadvantages
Reading speed	Number of pages above maximum or below minimum threshold	Accuracy prediction improves significantly	Mixed results for TP rate
Disengagement patterns	Separation of "disengaged-long" and "disengaged-fast" behaviour	Better TP rate for "disengaged-long" in most cases; works better in combination with reading speed attributes	Slightly lower accuracy rate than standard model due to additional class
Exclusion of exploratory sessions	Exclusion of the very first sequence after login when exploratory behaviour occurs	Better accuracy and TP rate for "disengaged-long" in most cases	Exploratory behaviour does not occur only at the first login; decrease of the TP rate for the "disengaged-fast" behaviour

The first refinement study showed that the general approach can be improved by using two additional attributes related to reading speed. These attributes are system-specific, i.e. the parameters would need to be computed for each particular system. Their impact seems to depend on the system as well. An overall increase of accuracy was observed for both HTML-Tutor and iHelp systems, with a greater increase for iHelp. While the TP rates increased for iHelp, the results were mixed for HTML-Tutor. Therefore, we concluded that for iHelp the benefit of using the reading attributes is twofold: considerably increased prediction and decreased time for processing. For HTML-Tutor, however, the results do not allow any conclusion and further investigation is required. One of the aspects that may account for the difference of results obtained with the two systems may lie in the way they are deployed; while the iHelp course was used in a formal educational setting, HTML-Tutor is freely available on the web. Although their structure and 'interaction possibilities' are very similar, the actual interactive behaviour may differ exactly because the first is somehow constrained while the latter is not. The constraints do not seem to be within the systems or the way the systems can be used, but in the goals the users have when using them. For example, on the one hand, when a system is used in educational settings, the learners may be more focused and systematic in their use of it, especially when the users are distance-learning students that usually have a job, as is the case of iHelp. On the other hand, when the system is freely available, it could be used by students in formal education as extra material and a way of testing their knowledge; it could be used by people interested in learning HTML without being in formal education, or by people interested only in a particular theme to remind them of concepts they have forgotten, etc. As information on the users' status was not available, we could not look into the different behaviour of these possible groups. Also, although HTML-

Tutor asks for the goals of the user when registering with the system, most users did not select one. Further investigation with data from a system that has such information available could bring some light on this matter.

Currently our approach uses average speed measurements for the reading speed attributes. Ideally, individual differences in reading would be accounted for, by having the reading speed measurements personalised to each user's reading rate. This would be easier to do in more constrained environments than in systems freely available on the web. A short test before using the system may provide this information, although some users, if not the majority, would skip it in a free environment.

The second refinement study showed that two patterns of disengagement can be distinguished: "disengaged-long" and "disengaged-fast". This distinction is valuable for personalized intervention. These two patterns of behaviour are in line with existing results described in the literature (cf. Section 2), but extend and generalize them: "blind guessing" in Beck (2005) or "unmotivated-guess" in Johns & Woolf (2006) are similar to the fast click through pages (denoted as "disengaged-fast" in our approach), as they describe students' rush and lack of attention (see Table 13). However, they are both specific for one type of learning activity, i.e. problem solving. We are not aware of an existing pattern that would correspond to the "disengaged-long" pattern found in our research. In systems where only problem solving activities are available, it is more likely to have a "disengaged-fast" pattern only, while in systems where both learning and problem-solving activities are present, it is more likely to have a "disengaged-long" as well as a "disengaged-fast" pattern. However, the fact that the interaction with the system for the two types of learning activities, i.e. reading and problem solving, is considerably different does not imply that the "disengaged-fast" behaviour can not occur during other learning activities. Therefore, further differentiation may be necessary for a more accurate prediction and a more personalised intervention. The opposite situation may occur as well – "disengaged-long" behaviour during problem solving activities, although the literature does not report on anything similar; in our data we also noticed that such behaviour is uncommon.

Table 13 Patterns of disengagement in our approach vs. related approaches.

Patterns	Our approach	Related approaches
Long time on a page	"Disengaged-long"	No correspondence
Click fast through pages	"Disengaged-fast"	Blindly guess (Beck, 2005)
		Unmotivated (guess/ hint) (Johns&Woolf, 2006)

Another difference that distinguishes our approach from related approaches concerns the domain: while existing results were based on rather technical domains, i.e. mathematics, we were able to demonstrate and validate our approach in a domain that is not as structured and hierarchical. We expect that the patterns observed may generalize more easily to many other domains, precisely because they are not highly dependent on structure.

The results of the patterns of disengagement study showed a slight decrease in the accuracy of the prediction which was expected due to the introduction of an additional class, but also a good distinction between the two patterns. Moreover, an increase of the true positive (TP) rate for the "disengaged-long" class is observed. This last aspect may be the most important one, as disengagement during learning activities (when the "disengagement-long" behaviour is more likely to occur) is potentially more harmful for learning than disengagement during test-type activities because in most on-line courses, the knowledge acquisition phase precedes and is prerequisite to a testing phase. Therefore, identifying the "disengaged-long" behaviour is the essential first step that allows personalized intervention.

Such intervention may include, but is not limited to: (a) investigate more on motivational aspects, e.g., engage the learner in a dialog to identify motivational characteristics (Cocea, 2006), (b) act upon known motivational characteristics of the learner, e.g., if the learner does not feel confident about the learned material, intervene either automatically or via the tutor with the aim of increasing their confidence (Hurley & Weibelzahl, 2007), (c) identify learners that may need help from the tutor, e.g., let the tutor prioritise among the disengaged learners, (d) provide help automatically to the tutor, e.g., provide priorities among disengaged learners depending on level of knowledge and motivational issues, leaving the final decision with the tutors, or (e) provide help automatically directly to the learner, e.g., give feedback on their performance accounting for their motivational characteristics and suggesting further learning activities.

The last study considered the particular case of what we called exploratory sequences. These are characterized by an exploratory behaviour and usually occur at the first login of a learner to the system. Given the fact that this behaviour is quite different and seems somehow "chaotic" compared to that observed in subsequent actions, we have explored their influence on the prediction values, expecting an improvement of prediction. While, unsurprisingly, we were able to demonstrate this effect for most cases, we observed a decrease in the prediction of the "disengaged-fast" (DF) pattern. This may be explained by the small number of DF instances and the fact that it dropped even more by the exclusion of exploratory sequences (mostly annotated as DF). Looking at the overall results for both HTML-Tutor and iHelp, the exclusion of these sequences has more benefits than drawbacks when judged strictly from the statistical point of view.

However, exploratory behaviour may occur not only at the first contact with a system, but within later interaction as well, without necessarily being a distraction from learning. In fact, the exploration may even facilitate better use of the system; for example, exploring its capabilities, like the glossary and statistics (of scores and material covered), allows quicker access to them when needed. While most users would explore the system at the first login, the exploratory behaviour varies in length and rigour. Some users prefer to find out as many capabilities of the system as possible before starting to use it while others start using the system after a minimal exploration and explore the system again when they need a certain capability that was not covered in their initial exploration. Given that the exploratory behaviour may occur not only when they first use the system but also at various points thereafter, eliminating *only* the first exploratory sequences from training may have a negative impact on the performance of the models as exploratory behaviour may not be classified correctly any more. This may explain the decrease in performance for some methods. As for the educational aspect of these exploratory sequences, even if a clear improvement in performance would occur, it is not obvious that this would be the way to proceed. This is an example of the difference between "classic" data mining and educational data mining, where the educational dimension should have prevalence over performance.

6. Conclusions

In this paper, we have presented three studies aiming to refine a general approach to disengagement prediction in e-Learning systems. This work addresses a limitation of current e-Learning systems: not taking into account the motivational characteristics of the learner. The studies presented in this paper are meant to improve the prediction of engagement levels, with a special focus on disengagement, as one aspect of (lack of) motivation. The goal is to be able to predict the level of engagement of the learner, which in turn would allow for personalized intervention based on both knowledge and motivational characteristics. The idea

is also to monitor the learner's behaviour in interacting with the system in an unobtrusive way so as not to interrupt the learning and to intervene only when necessary.

Our approach to detection of disengagement is simple and the information needed is related to actions that take place in most learning environments: reading pages and taking tests (solving problems). Therefore, this approach could be generalized to other systems, the validation study being an example for that. The similarity of results across data mining methods is also an indicator of the consistency of our approach. However, when integrating this approach in a real system, a decision must be made as to which prediction model to use. Among the range of possible procedures, the following two seem most advantageous in terms of time and space complexity, as well as scalability: (a) an initial test on data from the system followed by the usage of the average among the first three methods that perform best (majority vote); (b) an initial test on data from the system followed by the usage of the best performing method.

Besides this aspect, other courses of actions could be taken, like incrementally updating the information at some interval of time like, for example, every minute. Another possibility would be a survival analysis to detect when a learner is about to drop out. However, we think that having a trajectory of the learners' engagement status, with the three types of behaviour, i.e. engaged, "disengaged-long" and "disengaged-fast", allows a better diagnosis and could be of better use for the tutor or the system when deciding for a certain personalised intervention strategy.

One of the major challenges in our approach was to define disengagement in the context of e-Learning environments in terms of actions of learners when interacting with the system. This challenge was even more difficult due to the type of systems we wanted to look at – more specifically systems that provide learning content as well as problem-solving activities. Most research on motivation focused on problem solving activities which are more specific and hence, easier to assess and model in terms of disengagement compared to other learning activities. To overcome this problem we used human experts that assessed the level of engagement of learners based on their actions. This assessment was subsequently used in building the prediction models.

The studies presented in this paper brought more insight on the disengagement behaviour within e-Learning systems. The lessons learned include: (a) reading speed characteristics can help identify the levels of engagement; we found that simple attributes related to reading speed can improve the prediction level; however, this seems to be system-dependent or rather context-dependent: within an educational setting or freely on the Web; (b) two patterns of disengagement can be observed in the behaviour of learners: "disengaged-long" and "disengaged-fast"; the "disengaged-long" pattern is primarily associated with reading activities, while "disengaged-fast" occurs both during reading and problem solving; (c) a balance between educational benefits and prediction performance levels needs to be considered in educational data mining; in our particular case, although excluding the exploratory behaviour from the prediction model training improves the prediction, it would be wrong from an educational point of view to exclude this behaviour as if it were a distraction from the goal when, in fact, exploring the system may lead to a better use of what it offers to the learner.

Another challenge very much related to those mentioned previously is the subject domain. Most previous research used technical domains like mathematics or programming which are more "controllable" compared to non-technical domains. The domain used in our approach, HTML, is at the junction between technical and non-technical domains, therefore, allowing an easier generalization of our approach to other domains, including non-technical ones. However, the interaction design of the system may limit the generalisability of these findings; we looked at web-based systems that include reading and problem solving-activities and it is

unlikely that our specific findings would extend beyond this type of environments. Nevertheless, one lesson learned is that even if a domain is less structured or even structureless, it does not necessarily impair the possibility of modelling the user's activity. Still, the modelling process as presented here is of an explorative nature and, in this sense, closer to typical data-mining by aiming to discover information which is hidden in the data.

The approach presented here was specifically tailored to learning environments, but, considering the last observation, its applicability may stretch to other interactive behaviours as well. Voluntary and involuntary interruptions occur frequently when using computer or web-based systems. Involuntary interruptions, which would correspond to the "disengaged-long" pattern, are likely to decrease performance on more complex tasks (Speier et al., 2003). In such situations, a brief summary of what the user was previously doing may facilitate the "re-engagement" with the task. To the same purpose, spatial presentation formats could be used as they have been proven to mitigate the effects of interruptions, while symbolic formats have not (Speier et al., 2003). For example, an area where this could be useful is e-commerce. When a user has been inactive for some time, displaying the products previously viewed may be very helpful; depending on the previous activity level, the summary could be accompanied by product recommendations from the system.

In summary, we propose a simple approach for disengagement prediction that extends beyond previous approaches by including other learning activities besides problem solving. This approach gives very good results using attributes from only two actions: reading pages and taking tests, and can distinguish between two patterns of disengagement: "disengaged-long" and "disengaged-fast". Its simplicity and the characteristics of the chosen domain – HMTL – make it easier to generalize across systems and domains.

Acknowledgements

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, Christopher Brooks and Scott Bateman from University of Saskatchewan, Canada, for their generous support. We would also like to thank the reviewers for their many helpful comments and suggestions.

References

Arroyo, I. and Woolf, B.P.: 2005, '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*, Amsterdam: IOS Press, pp. 33-40.

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

Bandura, A.: 1986, 'Social foundations of thought and action: A social cognitive theory'. Englewood Cliffs, NJ: Prentice Hall.

Beal, C.R. and Lee, H.: 2005, 'Creating a pedagogical model that uses student self reports of motivation and mood to adapt ITS instruction'. *Proceedings of the Workshop on Emotion and Motivation in Educational Software (EMES)*, Amsterdam: IOS Press. Retrieved on 10/06/2006 from http://k12.usc.edu/Pubs/.

Beal, C.R., Qu, L. and Lee, H.: 2006, 'Classifying Learner Engagement through Integration of Multiple Data Sources'. *Proceedings of the 21st National Conference on Artificial Intelligence*. Menlo Park: AAAI Press. Retrieved on 10/06/2006 from http://k12.usc.edu/Pubs/.

Beck, J.: 2005, '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*, Amsterdam: IOS Press, pp. 88-95.

Becta, UK.: 2006, 'The benefits of an interactive whiteboard'. Retrieved on 11/06/2006 from http://schools.becta.org.uk/index.php?section=tl&catcode=ss_tl_use_02&rid=86.

Breiman, L.: 1996, 'Bagging predictors'. *Machine Learning* **24**(2), 123-140.

Chen, G.D., Shen, G.Y., Ou, K.L. and Liu, B.: 1998, 'Promoting motivation and eliminating disorientation for web based courses by a multi-user game'. Paper presented at *The ED-MEDIA/ED-TELECOM 98 World Conference on Educational Multimedia and Hypermedia and World conference on Educational Telecommunications*, June 20-25, Germany.

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

Cocea, M. and Weibelzahl, S.: 2007a, 'Eliciting motivation knowledge from log files towards motivation diagnosis for Adaptive Systems'. In: C. Conati, K. McCoy, and G. Paliouras (eds.): *Proceedings of The 11th International Conference on User Modelling*, Lecture Notes in Artificial Intelligence (LNAI), Springer, Berlin, vol. 4511, pp. 197-206.

Cocea, M. and Weibelzahl, S.: 2007b, 'Cross-System Validation of Engagement Prediction from Log Files'. In: Duval, E., Klamma, R. and Wolpers, M. (eds.): *Creating New Learning Experiences on a Global Scale, Second European Conference on Technology Enhanced Learning, EC-TEL 2007*, Lecture Notes in Computer Science (LNCS) , Springer Berlin/ Heidelberg, vol. 4753, pp. 14-25.

Cocea, M.: 2006, 'Assessment of motivation in online learning environments'. In: V. Wade, H. Ashman and B. Smith (eds.): *Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer-Verlag, pp. 414-418.

Connolly, T. and Stansfield, M.: 2006, 'Using Games-Based eLearning Technologies in Overcoming Difficulties in Teaching Information Systems'. *Journal of Information Technology Education* **5**, 459-476.

Csikszentmihalyi, M.: 1997, 'Finding Flow: The Psychology of Engagement with Everyday Life', BasicBooks, New York.

De Vicente, A. and Pain, H.: 2002, '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*. Berlin: Springer-Verlag., pp. 933-943.

Farzan, R. and Brusilovsky, P.: 2005, '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.

Gussak, D. and Baylor, A. L.: 2003, 'Constructing Agents for Self-Learning: Animated Agents as Expressive Vehicles'. In: D. Lassner and C. McNaught (eds.): *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Chesapeake, VA: AACE, pp. 477-478.

Hurley, T. and Weibelzahl, S.: 2007, 'Eliciting adaptation knowledge from on-line tutors to increase motivation'. In: C. Conati, K. McCoy, K. and G. Paliouras (eds.): *Proceedings of The 11th International Conference on User Modelling*, Lecture Notes in Artificial Intelligence (LNAI), Springer, Berlin, vol. 4511, pp. 370-374.

Ishii, T., Saitou, M. and Hiramoto, S.: 2004, 'An instructional design for developing learning contents by subject matter experts'. *Proceedings of the 9th World Conference on Continuing Engineering Education*, pp. 469-474.

Johns. J. and Woolf, B.: 2006, 'A Dynamic Mixture Model to Detect Student Motivation and Proficiency'. *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, pp. 163-168.

Keller, J. M.: 1987, 'Development and use of the ARCS model of instructional design'. *Journal of Instructional Development* **10**(3), 2-10.

Lombard, M., Snyder-Duch, J. and Campanella Bracken, C.: 2003, 'Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research'. Retrieved on 11/06/2007 from http://www.temple.edu/mmc/reliability.

Machado, I., Martins, A. and Paiva, A.: 1999, 'One for all and all for one: a learner modelling server in a multi-agent platform'. *Proceedings of the Seventh International Conference on User Modelling*, Springer, Wien, pp. 211-221.

Martyn, M.: 2007, 'Clickers in the classroom: An active learning approach'. *EDUCAUSE Quarterly* **30**(2), 71-74.

Mitchell, T. M.: 1997, 'Machine Learning', McGraw-Hill.

Pintrich, P. and Schunk, D.: 2002, 'Motivation in Education: Theory, Research and Applications', 2nd edition, Prentice-Hall, Englewood Cliffs, NJ.

Qu, L., Wang N. and Johnson, W. L.: 2005, 'Detecting the Learner's Motivational States in an Interactive Learning Environment'. In: C.-K. Looi, G. McCalla, B. Bredeweg and J. Breuker (eds): *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, IOS Press, pp. 547-554.

Quinlan, R.: 1993, 'C4.5: Programs for Machine Learning', Morgan Kaufmann Publishers, San Mateo, CA.
Shneiderman, B., Alavi, M., Norman, K. and Borkowski, E.: 1995, 'Windows of opportunity in electronic classrooms'. *Communications of the ACM* **38**(11), 19-24.
Speier, C., Vessey, I. and Valacich, J.S.: 2003, 'Effects of Interruptions, Task Complexity, and Information Presentation on Computer-Supported Decision-Making Performance'. *Decision Sciences* **34** (4), 771-797.
Walonoski, J. and Heffernan, N. T.: 2006a, 'Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems'. In: Ikeda, Ashley & Chan (eds.): *Proceedings of the 8th International Conference in Intelligent Tutoring Systems*. Springer-Verlag: Berlin, pp. 382-391.
Walonoski, J. and Heffernan, N. T.: 2006b, 'Prevention of Off-Task Gaming Behaviour within Intelligent Tutoring Systems'. In: Ikeda, Ashley & Chan (eds.): *Proceedings of the 8th International Conference in Intelligent Tutoring Systems*. Springer-Verlag: Berlin, pp. 722-724.
Weber, G., Kuhl, H.-C. and Weibelzahl, S.: 2001, 'Developing adaptive internet based courses with the authoring system NetCoach'. In *Hypermedia: Openness, Structural Awareness, and Adaptivity* (LNAI 2266), Springer, Berlin, pp. 226-238.
Witten, I.H. and Frank, E.: 2005, 'Data mining. Practical Machine Learning Tools and Techniques'. Second Edition, Morgan Kauffman Publishers, Elsevier.
Zhang, G., Cheng, Z., He, A. and Huang, T: 2003, 'A WWW-based Learner's Learning Motivation Detecting System'. *Proceedings of International Workshop on "Research Directions and Challenge Problems in Advanced Information Systems Engineering"*, Honjo City, Japan, September 16-19, http://www.akita-pu.ac.jp/system/KEST2003/.

Authors' vitae

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 "Al. I.Cuza" University of Iasi 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 London Knowledge Lab, School of Computer Science and Information Systems, Birkbeck College, University of London. Her research interests include intelligent learning environments, educational data mining, user modelling 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.