Performance Prediction of Elementary School Students in Search Tasks

Roberto González-Ibañez[a], Luz Chourio-Acevedo[a,b] and María Escobar-Macaya[a]

[a]Universidad de Santiago de Chile, Avenida Libertador Bernardo O'Higgins n° 3363. Estación Central, Santiago, Chile
[b]Centro Nacional de Desarrollo e Investigación en Tecnologías Libres, Avenida Humberto Carnevalli, Edificio CENDITEL, Mérida, Venezuela

Abstract

In the last two decades, the use of online resources in educational settings has seen an unprecedented growth. Regrettably, students' online inquiry competences (OIC) are not necessarily well developed to face problems involving information intensive domains. While different OIC development approaches have been proposed to address this situation, these fail in timely identifying their effects on students' OIC applied to practical search scenarios. To address this drawback, in this article we study models to predict students' search performance in the context of an OIC evaluation test. Our approach focuses on exploiting demographic, behavioral, cognitive, and affective features, to predict – at four points of the overall search process – whether students succeed or fail in finding relevant documents to accomplish a research task. Our preliminary results show that it is possible to anticipate the overall search performance of students with moderate accuracy at the 25%, 50%, 75%, and 90% of the search session progress. These findings illustrate potential benefits and limitations of using non-obstrusive aggregated signals to timely predict search performance in learning contexts.

Keywords

Search perfomance, prediction, classification, elementary school

1. Introduction

Internet, and particularly the World Wide Web (WWW), has become the main resource for students who look for information to complete their school assignments. Although abundant, not all the content on the Web is curated[1]. This poses a major problem for students who may not be well equipped in terms of OIC. Indeed, knowing what information is needed and how to search for it (i.e., some component skills of OIC) is crucial to succeed in online research [2]. To tackle this problem, different approaches to help students in the development of OIC have been proposed [1, 3]. A fundamental limitation of these approaches is their inability to timely determine whether students will succeed or fail when engaging in actual search tasks.

In the context of OIC development, knowing in advance how a student will perform in a search task could be particularly useful to both educators and students. First, educators could offer opportune feedback and support to their students, thus avoiding late evaluations typically available only after tests are completed. Second, students themselves could be more aware of their own performance, which could help them to correct themselves or look for support. In educational contexts, prediction focuses on forecasting performance by estimating unknown values of variables that characterize students. Such values typically relate to performance, knowledge, and scores. Prediction can be also used to: identify learning styles, determine whether a student will answer a question correctly, model knowledge changes, and determine non-observable learning variables [4].

In this article, we explore the possibility to anticipate student's search performance by exploiting a set of demographic, behavioral, cognitive, and affective features through machine learning. The remaining sections of this article are organized as follows. First, we describe the methodological approach adopted for this work. Second, we present preliminary results. Finally, we conclude with a discussion of the results, their implications, and future work.

2. Method

2.1. Dataset

To conduct this study, we relied on a subset of the data collected as part of the iFuCo project [5]. Our sample contains search sessions from 350 Finnish students performing two independent research tasks, this in the context of an evaluation of OIC. A summary of demographic data of the students whose records are included in our study is presented in Table1.

Records in this dataset were captured through NEURONE (oNlinE inquiry expeRimentatiON systEm) [6].

Proceedings of the CIKM 2020 Workshops, October 19-20, 2020, Galway, Ireland
EMAIL: roberto.gonzalez.i@usach.cl (R. González-Ibañez); luz.chourio@usach.cl (L. Chourio-Acevedo); maria.escobarm@usach.cl (M. Escobar-Macaya)

Table 1
Demographic data of the students

Finnish cities	Tampere, Jyväskylä, Turku
Grades	Fifth and sixth
Ages	12-13 years old
Girls	48.18%
Boys	51.82%

This system offered a realistic simulation of a search engine operating on a controlled collection of web documents for each research task. The document collection was developed by the research team and comprised 20 web pages per tasks, three of them defined as relevant. Regarding the latter documents, these were created by researchers and all three were required to be found in order to accomplish each research task.

The dataset contains various types of data, which includes behavioral, cognitive, affective, and demographic variables. Table 2 lists all the variables included in this dataset.

2.2. Analysis procedure

Our general approach to evaluate the feasibility of predicting search performance focuses on four moments within students' search sessions: early (25%), middle (50%), late (75%), and close-to-end (90%). Based on this nominal division, we aim to compare different models in the classification task of whether students will fail or succeed in the overall search task (i.e., binary classification).

To determine whether a student failed or succeeded in the search tasks, we relied on search score, a process-based measure defined in [7]. This measure accounts for both, the success in finding relevant documents and mistakes made during the search process. Since search scores range from 0 to 5, we defined a threshold of 3.3 to balance the data. This value was set to keep a slightly balanced dataset of pass/fail cases. Thus, students with a score of 3.3 or higher were labeled as Pass (46%), whereas those below this threshold were labeled as Fail (54%).

Following, we normalized search sessions, which lasted a maximum of 8 minutes. Normalization was necessary to have all sessions in a common duration scale, which were now expressed from 0% to 100%. Next, we proceeded to generate four additional subsets of sessions based on the four moments stated above. As a result, the first set contains session data of each student from 0% to 25%, the second set comprised data from 0% to 50%, and so forth. Each subset contained the Pass or Fail label computed at 100% of each search session (See Figure 1).

Table 2
Dataset attributes

Attribute	Description
Behavior (during the session)	
Total.Time (TT)	Segment total time
Stay.Pag.Relv (SR)	Dwell time in relevant pages
Stay.Pag.NonRelv(SnR)	Dwell time in non-relevant pages
Query.Time (QT)	Query writing time
Count.Queries (CQ)	Number of queries
Q.Mod (QM)	Number of query modifications
Q.Entropy (QE)	Average query entropy
Total.Cover (TC)	Total coverage
Usf.Cover (UC)	Useful coverage (dwell time ≥ 30 seconds)
Relv.Coverage (RC)	Number of relevant pages visited
Clicks.Relv (CR)	Number of clicks within relevant pages
Clicks.NonRelv (CnR)	Number of clicks within non-relevant pages
Mouse.Mov.Relv (MR)	Number of mouse movements within relevant pages
Mouse.Mov.NonRelv(MnR)	Number of mouse movements within non-relevant pages
Scroll.Mov.Relv(SMR)	Number of scrolls within relevant pages
Scroll.Mov.NonRelv(SMnR)	Number of scrolls within non-relevant pages
Demographic	
Sex	Girl, Boy
Affective (SAM-based scale [8])	
Pos	Valence (Positive - Negative scale)
Cal	Activation (Calm - excited scale)
Cognitive(Survey)	
Prior.Knowledge (PK)	Prior knowledge on task topic (1 to 5 scale)
Perceived.Difficulty (PD)	Perceived task difficulty level (1 to 5 scale)
class	Pass (A), Fail (R)

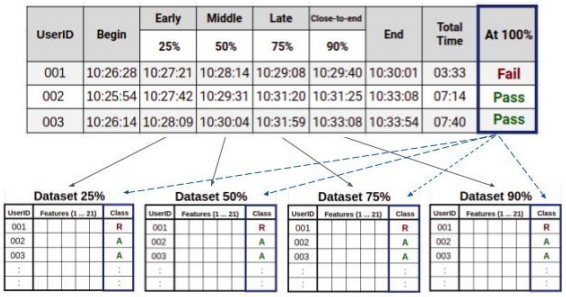

Figure 1: Subset generation based on normalized search sessions.

We followed the Knowledge Discovery in Data bases (KDD) process with each dataset, thus we performed data selection, preprocessing, transformation, data mining, and evaluation/interpretation to derive knowledge. To implement these stages, we used both Weka and R.

After preprocessing data, we ended up with a total of 660 full search sessions. For the purpose of this study, we discarded incomplete sessions (due to con-

Table 3
Automatic attribute evaluation.

	CFSSubsetEval	InfoGainAttributeEval
25%	TT, SnR, QE, TC, MR	TT, SnR, MR, QE, TC, Sex
50%	TT, SR, SnR, QE, TC, MR, MnR	TT, SnR, MR, SR, MnR, QE, SR, TC, RC, Sex, UC
75%	TT, SR, SnR, RC, MR	RC, MR, SnR, TT, SR, TC, SMR, MnR, Sex
90%	TT, SR, SnR, RC, MR	RC, SR, MR, SnR, TT, TC, SMR, MnR, Sex

nection problems) and those with corrupted data. These problems were mainly caused by connection problems or incompatibility of browsers with NEURONE.

Once features were selected, preprocessed, and transformed, we created vectors of features containing aggregated session data (mostly behavioral) until the corresponding interval (i.e., 25%, 50%, 75%, 90%). In addition, these vectors contained prior-session features from demographic, cognitive, and affective variables. Finally, Pass/Fail labels (i.e., class) were added. Overall, our vectors contained 21 features plus the class.

With these vectors, we proceeded to identify prominent features and build binary classifiers through different algorithms and approaches. Results achieved by these classifier in the task of determining the pass/fail labels are presented in the following section.

3. Results

After building vectors in each subset, we ran automatic attribute evaluation in order to determine which features could contribute the most to the classification task. This procedure was conducted using two Weka algorithms, namely, CFSSubsetEval and InfoGainAttributeEval. As a result of this procedure, eight groups of features were identified, two per subset, as shown in Table 3. Additionally, we performed attribute scanning, which led us to discard or include other features in all four subsets. On the one hand we discarded variables related to clicks in relevant and non-relevant pages since they did not improve nor worsen classification performance. In other words, their presence increased problem dimensionality in terms of features unnecessarily. On the other hand, we included cognitive measures (i.e., prior knowledge and perceived task difficulty) and an affective measure (Pos) as input variables to the search process [9].

Next, by combining the selected features (those in Table 3 and positivity score (Pos)) following a brute-force approach, we built classifiers through linear regression, logistic regression, Naïve Bayes, JRIP, J48, random forest, multilayer perceptron, SMO RBF kernel, and SMO poly kernel. All models were trained and tested through 10-fold cross-validation. The classes in all cases were linked to the Pass/Fail labels computed at 100%, hence our classifiers were actually prediction models attempting to determine the overall search performance of students. Results were compared in terms of precision, F-Measure, number of attributes, and area under the ROC curve (AUC). A summary of the best results achieved at each time point (in terms of AUC) is presented in Table 4.

Table 4
Support metrics of the best models obtained(class=Pass/Fail).

	25%	50%	75%	90%
Model	Classification via Regression	Classification via Regression	Random Forest	Logistic Regression
# Features	11	10	6	10
Features	TT, SR, TC, RC, UC, QM, QE, SMR, MnR, Sex, Pos	TT, SR, TC, RC, UC, CQ, QM, Sex, PK, PD	TT, SnR, TC, RC, MR, Sex	SnR, TC, RC, UC, QT, CQ, QE, MR,SMR, SMnR
Area under curve ROC	0.736	0.770	0.827	0.866
Error (%e)	30.00%	27.28%	23.64%	19.55%
Precision	0.690	0.734	0.760	0.792
F-Measure	0.669	0.691	0.783	0.790

4. Discussion

As illustrated in Table 4, different models, with different set of features achieved the highest AUC at different time points. At an early stage of students' search processes (i,e., 25%), our best model is based on linear regression over 11 features with an AUC of 0.736 and an error of 30%. Then, at 50% of search sessions, the best model is also based on linear regression, however the set of features is slightly different and performance increases in 4.6% in terms of AUC. Later on, at 75% of search progress, the best model is based on random forest over six features. In this case, performance in terms of AUC shows an increment of 12.36% with respect to our early-stage best model. Also, a reduction in error by almost 7% is noted. Finally, very late at students' search sessions (i.e., 90%), the best model is based on logistic regression over 10 features. In this case, AUC is 0.866, whereas error was reduced to 19.55%.

In this group there are features involving time spent in relevant and non-relevant pages, query-related features, document coverage, and mouse movements, to name a few. In addition, we highlight that sex (i.e., a demographic feature) appears as a prominent feature used by our best performing models at 25%, 50%, and 75%. Additionally, an affective feature (Pos, which ex-

press valence in a negative-positive scale) was present in the best performing model at 25%. Likewise, prior knowledge on the topic (PK) and perceived task difficulty (PD) are used in the best performing model at 50%. We note that these particular input features, which are captured before search sessions start, seem to play some role in the way search processes are carried out. On the one hand, the fact that sex appears in three out of four models (Table 4), indicates that girls and boys may exhibit particular search patterns that could be linked to search performance. On the other hand, the presence of an affective feature (i.e., Pos) also supports the idea that searchers' initial affective states may shape their search behaviors and their relevance assessments (e.g., participants in negative states being more systematic than those in positive states) [10, 9].

As expected, the earlier in the search process, the higher the level of uncertainty to correctly predict the overall search performance. On the contrary, the later in the search process, the higher the level of certainty to determine whether students will succeed or fail once search sessions were completed. Despite the low-performance of classification models at 25%, this shed light that, to some extent, it is possible to timely predict students' search performance. More interestingly, our best model is rather simple and it relies on variables that can be captured easily in controlled and open environments (e.g., mouse actions, query formulation features, some demographic data).

As for limitations of our prediction approach, the fact it is based on aggregated data at different moments of students' search leads to data loss. Indeed, the history of students' actions while searching for information (e.g., query formulation, page visit, scrolling actions, query reformulation, bookmarking, etc.) is compressed into single measures (e.g., means, sums, counts). Such chain of actions could be crucial to anticipate how students will perform in the short and long term. In this sense, our future work will concentrate in studying prediction approaches that take into account the dynamics of search behaviors. Among these approaches we consider Markovian models and SVM with string-based kernels.

4.0.1. Acknowledgment

The work described in this article was partially supported by the TUTELAGE project funded by the National Agency for Research and Development (ANID) (FONDECYT Regular, grant no. 1201610); the Vicerrectoría de Postgrado of the Universidad de Santiago de Chile; and the iFuCo project funded by the Academy of Finland (grant no. 294186) and ANID (grant no. AKA/EDU-03).

References

[1] F. Baji, Z. Bigdeli, A. Parsa, C. Haeusler, Developing information literacy skills of the 6th grade students using the big 6 model, Malaysian Journal of Library & Information Science 23 (2018) 1–15.

[2] S. Majid, S. Foo, Y. Chang, Appraising information literacy skills of students in singapore, Aslib Journal of Information Management (2020).

[3] H. Zhang, C. Zhu, A study of digital media literacy of the 5th and 6th grade primary students in beijing, The Asia-Pacific Education Researcher 25 (2016) 579–592.

[4] C. Romero, S. Ventura, Educational data mining: a review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40 (2010) 601–618.

[5] M. Mikkila-Erdmann, E. Sormunen, T. Mikkonen, N. Erdmann, C. Kiili, M. Quintanilla, R. González-Ibáñez, P. Leppanen, M. Vauras, A comparative study on learning and teaching online inquiry skills in finland and chile, in: European Conference on Information Literacy (ECIL), volume 18, 2017, p. 2017.

[6] R. González-Ibáñez, D. Gacitúa, E. Sormunen, C. Kiili, Neurone: online inquiry experimentation system, Proceedings of the Association for Information Science and Technology 54 (2017) 687–689.

[7] E. Sormunen, R. González-Ibáñez, C. Kiili, P. H. Leppänen, M. Mikkilä-Erdmann, N. Erdmann, M. Escobar-Macaya, A performance-based test for assessing students' online inquiry competences in schools, in: European Conference on Information Literacy, Springer, 2017, pp. 673–682.

[8] M. Bradley, P. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.

[9] R. González-Ibáñez, C. Shah, Performance effects of positive and negative affective states in a collaborative information seeking task, in: CYTED-RITOS International Workshop on Groupware, Springer, 2014, pp. 153–168.

[10] R. Sinclair, M. Mark, The effects of mood state on judgemental accuracy: Processing strategy as a mechanism, Cognition & Emotion 9 (1995) 417–438.