Prediction of Perceived Disorientation in Online Learning Environment with Random Forest Regression

I. AKÇAPINAR
Hacettepe University, Turkey
II. COŞGUN
Hacettepe University, Turkey
AND
III. ALTUN
Hacettepe University, Turkey

In this study, it is aimed to predict users' perceived disorientation by using a data mining technique: Random Forest Regression (RFR). Two RFR models are designed to predict users' disorientation scores. The models were generalized with 10-fold Cross Validation (CV). In the first model, log based metrics were used as explanatory variables. In the second, log based metrics, eye metrics and self-report metrics were included in the model. According to findings, our RFR models predict perceived disorientation score with high accuracy. First model's R^2 is 57.8 %, second model's R^2 is 63.5 %. These results showed that adding eye metrics and self-report metrics to the model increased the predictive performance.

Key Words and Phrases: Disorientation, random forest, data mining, navigation log, eye tracking.

1. INTRODUCTION

In a hypermedia environment, information is presented nonlinearly and connectively unlike from written text. While this flexible structure provides learners with great browsing freedom, still there is a risk for some learners to become "lost in hyperspace" (Botafogo, Rivlin, & Shneiderman, 1992). Getting disoriented is one of the major difficulties that users experience while navigating (Conklin, 1987).

1.1 Disorientation

Conklin (1987) defined disorientation as the tendency to lose one's sense of location and direction in a non-linear document. Disorientation, or the tendency to lose one's sense of location in a Web site, can cause users to become frustrated, lose interest, and experience a measurable decline in efficiency (McDonald & Stevenson, 1998). While disoriented users feel more cognitive load, their learning performance is also affected negatively (Madrida, Oostendorpb, & Melguizob, 2009). For those reasons, much hypermedia research has been devoted to this matter (Otter & Johnson, 2000). There are two major approaches to measure user disorientation levels: objective or subjective measurements (Gwizdka & Spence , 2007). Disorientation in objective measurements is usually addressed with the term "lostness".

Smith (1996) proposed an objective measure of lostness based on the ratios of visited and optimal node counts. Otter and Johnson (2000) added a link's weight to Smith's

Authors' addresses: G. Akçapınar, Department of Computer Education and Instructional Technologies, Hacettepe University, Ankara, Turkey. E-mail: gokhana@hacettepe.edu.tr; E. Coşgun, Department of Biostatistics, Hacettepe University, Ankara, Turkey; E-mail : erdal.cosgun@hacettepe.edu.tr; A. Altun, Department of Computer Education and Instructional Technologies, Hacettepe University, Ankara, Turkey. E-mail: altunar@hacettepe.edu.tr

formula, proposing a weighted lostness formula. They also proposed another measure which is concerned with the accuracy of users' mental models of websites. Dias and Sousa (1997) introduced the orientation ratio which measures the degree of disorientation as the indicator of the accuracy of information retrieval.

On the other hand, Ahuja and Webster (2001) claimed that a better way to assess users' disorientation is asking users directly how they feel after navigating in a website. For this purpose, the authors validated a questionnaire on self-perceived disorientation and found that their questionnaire predicted performance on web information retrieval tasks better than user actions.

Although disorientation is a common problem for the internet users, it is difficult to measure it (Herder E. , 2003). In this study, we combined different kinds of metrics including eye movement data to determine the most important predictors of disorientation with the help of a data mining technique: Random Forest Regression (RFR).

1.2 Random Forest Regression

RFR is an effective nonparametric statistical technique for high-dimensional analysis. Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Cosgun, Limdi, & W. Duarte1, 2011). The tree methods exhaustively break down cases into a branched, tree-like form until the splitting of the data is statistically meaningful, with unnecessary branches pruned using other test cases to avoid over-fitting (Choi & Lee, 2010). The generalization error for forests converges to a limit as the number of trees in the forest becomes large, and depends on the strength of the individual trees in the forest and the correlation between them (Breiman, 2001). Each tree in the forest is grown with following steps:

1. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
2. Grow a "random" tree, where at each node, the best split is chosen among randomly selected variables (m). The tree is grown to maximum size and not pruned back.
3. Use the tree to predict out-of-bag data.
4. Use the predictions on out-of-bag data to form majority votes.
5. Repeat, N times and collected an ensemble of N trees. Prediction of test data is done by majority votes from predictions from the ensemble of trees (Emir & Cabrera, 2009).

1.2.1 Variable importance

RFR determines the relative importance of each variable, through various methods, such as the calculation of the Gini Index, which assesses the importance of the variable and carries out accurate variable selection (Torri, Beretta, Ranghetti, Granucci, Ricciardi-Castagnoli, 2010). In every tree grown in the forest, put down the out-of-bag (oob) cases and count the number of votes cast for the correct class. After that randomly permute the values of variable m in the oob cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-m-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable m (Breiman & Cutler, 2004).

2. METHOD

Thirty prospective high school computer teachers from Computer Education and Instructional Technologies Department of Hacettepe University (8 females and 22 males) participated in this study.

Data was collected on a web site that was designed by one of the researchers with the purpose of learning a given task. The content of the task included the basic SQL queries (Select, Update, Delete, Insert) and their use within the PHP language. The web site was exploratory in nature and hyperlinked across concepts. This topic was selected because it was new information for the participating students. They have already learned basic PHP language. A networked structure was used to present information consisting of 22 web pages (about 1000 words) and 57 cross-reference links between pages.

Table 1 shows the descriptions of metrics. Students' perceived disorientation was measured using the Turkish version of the Ahuja and Webster's (2001) self-report instrument. This instrument adapted to Turkish by Cangöz and Altun (2010). There were 10 statements and subjects were asked to indicate whether they agreed or disagreed with the statements on a 5-point Likert scale. Disorientation scores were calculated by the sum of user ratings of these statements. Log metrics were gathered through log data, which were collected from ASP (Active server page) coded learning environment. Time spent on a web site (duration), revisited page counts (revisited), unvisited page counts (unvisited) and unique page counts (unique) were extracted from these log files. Eye movement data were collected by Tobii T120 eye tracker. Fixation count and fixation duration metrics were selected for the study. Students' exam scores related to PHP language are taken as the prior knowledge score.

Table I. Description of metrics

	Description
Eye metrics	
FDonTitle	Total fixation duration on titles
FConTitle	Total fixation count on titles
FDonContent	Total fixation duration on contents
FConContent	Total fixation count on contents
FDonSample	Total fixation duration on code samples
FConSample	Total fixation count on code samples
FDonLink	Total fixation duration on links
FConLink	Total fixation count on links
Log metrics	
Duration	Total time spent while performing learning task
UniqueNav	Total number of unique web pages
Revisited	Total number of revisited web pages
Unvisited	Total number of unvisited web pages
Self-report metric	
PK	Learners' prior knowledge tests score

2.1 Design and Procedure

All of the learners performed a learning task in network structured hypermedia. While they were performing the task, their navigation data were logged and their eye movements were recorded by the eye tracker device. They were given a maximum 10 minutes to study the material. After they completed the task, they were requested to complete the perceived disorientation scale.

Random Forest Regression (RFR), were implemented using R-2.11.0 which is an open source software. We have used the Random Forest package for RFR, the ModelMap package for data manipulation.

3. RESULTS

Two different RFR regression models were developed to predict users' perceived disorientation. The first model included users' navigation log data. For the second model, users' eye metrics and prior knowledge scores were added to the model. Users' perceived disorientation was the dependent variable in both of the models.

According to findings related to Model 1, time spent on a web site, revisited page count and unvisited page count were the most important predictor of perceived disorientation (Table II). First model's R^2 is 57.8 %.

According to findings related to Model 2, fixation duration on samples, fixation count on contents, time spent on a web site, fixation count on samples, prior knowledge, unique navigation and fixation duration on links were more important predictors of perceived disorientation among others (Table II). Second model's R^2 is 63.5 %.

Table II. Predictor importance for Model 1 and Model 2

	Importance
Model 1	
Duration	1.00*
Revisited	0.50*
Unvisited	0.49*
UniqueNav	0.40
Model 2	
FDonSample	1.00*
FConContent	0.93*
Duration	0.93*
FConSample	0.76*
PK	0.68*
UniqueNav	0.63*
FDonLink	0.52*
FDonContent	0.48
FConTitle	0.48
Revisited	0.40
FConLink	0.38
FDonTitle	0.38
Unvisited	0.35

Important variables marked with asterisks (*)

4. CONCLUSION

In this study, two different RFR models were proposed, and these models found to be a predictor of perceived disorientation with high accuracy. In eye movement research, when users produce more fixations on certain fields on the screen, their attention is thought to be on that field rather than the other areas on screen (Poole, Ball, & Phillips, 2004). This could indicate that the viewer is either having difficulty in extracting information or the object is more engaging in some way (Just & Carpenter, 1976). When eye metrics were included in the model, predictive performance was increased, indicating that eye metrics are important predictors for perceived disorientation. If researchers who

study disorientation have a chance to collect data with eye tracker, they must use it; otherwise, log based metrics can also be applied with reported accuracy.

Unvisited and revisited page counts were found to be good predictors in a log based model. When other metrics (prior knowledge and eye metrics) were included in the model, their (unvisited and revisited page counts) importance decreased. Herder (2003) also reported that no correlation was found between users' perceived disorientation and percentage of revisits. This finding can be interpreted as users may use revisits as a navigation strategy.

Time spent on a specific web page is an important predictor of disorientation in both models. Amadieu, Gog, Paas, Tricot, and Mariné (2009) found that low prior knowledge learners experienced higher disorientation than their higher counterparts during learning with a network structured concept map. In both models, prior knowledge was found to be an important predictor of user perceived disorientation.

We have a limitation on sample sizes. Because, in nature of these kind of educational data, number of users are limited. But data mining methods are very useful for determine the best solutions. In this study we have tried to implement Random Forest Regression algorithm for finding best model on "disorientation data". For discard the disadvantages of sample size, our entire process was contained within a 10-fold cross validation structure. This approach is helpful for generalized our findings. In future studies we are going to use this "base model" and try to increase our prediction performance.

REFERENCES

Ahuja, J., & Webster, J. (2001). Perceived disorientation: an examination of a new measure to assess web design effectiveness. *Interacting with Computers, 14*, 15-29.

Amadieu, F., Gog, T. V., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction*(19), 376-386.

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems, 10*, 142-180.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5–32.

Breiman, L., & Cutler, A. (2004). *Random Forests*. Retrieved from Random Forests: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp

Cangöz, B., & Altun, A. (2010). Bilgisayar geziniminde örtük bellek ve algılanan oryantasyon kaybının rolü. (The role of implicit memory and perceived disorientation in navigation). *16. Ulusal Psikoloji Kongresi* . Mersin: Mersin Üniversitesi.

Choi, M., & Lee, G. (2010). Decision tree for selecting retaining wall systems based on logistic regression analysis. *Automation in Construction*, 917–928.

Conklin, J. (1987). Hypertext: An introduction and survey. *IEEE Computer, 20*(7), 17-41.

Cosgun, E., Limdi, N., & W. Duarte1, C. (2011). High Dimensional Pharmacogenetic Prediction of a Continuous Trait using Machine Learning Techniques with Application to Warfarin Dose Prediction in African Americans. *Bioinformatics Advance Access*, 1-6.

Dias, P., & Sousa, A. P. (1997). Understanding navigation and disorientation in hypermedia learning environments. *Journal of Educational Multimedia and Hypermedia*, 173-185.

Emir, B., & Cabrera, J. (2009, May 10). Course Notes of "Exploring/Data Mining Pharmaceutical Data". İstanbul, Turkey.

Gwizdka, J., & Spence , I. (2007). Implicit measures of lostness and success in web navigation. *Interacting with Computers., 19*(3), 357-369.

Herder, E. (2003). Revisitation Patterns and Disorientation. *In Proceedings of the German Workshop on Adaptivity and User Modeling in Interactive Systems ABIS 2003* (pp. 291-294). Karlsruhe: University of Karlsruhe.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*, 441-480.

Madrida, R., Oostendorpb, H., & Melguizob, M. (2009). The effects of the number of links and navigation support on cognitive load and learning with hypertext: The mediating role of reading order. *Computers in Human Behavior*, 66-75.

McDonald, S., & Stevenson, R. J. (1998). Effects of text structure and priorknowledge of the learner on navigation in hypertext. *Human Factors*, 18–27.

Poole, A., Ball, L. J., & Phillips, P. (2004). In search of salience: A response time and eye movement analysis of bookmark recognition. In P. M. S. Fincher, *People and Computers XVIII-Design for Life: Proceedings of HCI 2004* (pp. 363-378). London: Springer-Verlag.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers*, 365-38.