Detecting Game Player Goals with Log Data

2014 Annual Meeting of the American Educational Research Association
Philadelphia, PA

Kristen E. DiCerbo

April 2014

Abstract

As gaming researchers attempt to make inferences about player knowledge, skills, and attributes from their actions in open-ended gaming environments, understanding game players' goals can help provide an interpretive lens for those actions. Games are generally far more open-ended than traditional assessments and as such, inferential evidence must take into account the context of the player's actions. Algorithms can help researchers identify particular goals in large log files of detailed player actions. This research uses Classification and Regression Tree methodology to develop and then cross-validate features of game play and related rules through which player behavior can be used to classify players according to their goals.

Keywords: Games, goal detections, assessment, log files

Detecting Game Player Goals with Log Data

Many recently-developed online learning environments provide open spaces for students to explore. These environments allow students to engage in a variety of behaviors within the environment so that they may develop or demonstrate knowledge, skills, and attributes (Code, Clark-Midura, Zap, & Dede, 2012; DiCerbo & Behrens, 2012). At the same time, there is growing interest in stealth assessment (Shute, 2011), or the use of data resulting from students' every day interactions in a digital environment to make inferences about player characteristics. For example, the novelty of players' solutions in the commercial *World of Goo* game was used to build models of player creativity (Shute & Kim, 2011), repeated attempts after failure to build models of persistence (DiCerbo, 2014; Ventura & Shute, 2012), and actions to reduce pollutions and increase jobs in SimCityEDU to assess systems thinking (Mislevy et al, 2014).

This use of data from natural activity in open-ended environments presents a challenge for interpretation. Much of the evidence we wish to use to assess skill proficiency and player attributes assumes that individuals are working towards the goal of completion of sub-tasks or levels within a game. However, research by Bartle (1996) on Multi-User Dungeons characterized four different types of player goals: achievement (reaching game goals), exploration (finding out as much as they can about how the virtual world works), socializing (interacting with other players), and imposition on others (disrupting play of others). These various goals provide different lenses for interpreting player behavior based on data in log files for games. For example, Sabourin, Rowe, Mott, & Lester (2011) describe off-task behavior including spending too much time in a location irrelevant to the task. However, if a student's goal is exploration of the entire space, this activity may not be irrelevant to their perceived task, and perhaps should be classified as on-task behavior. Even if the goal of a game is clearly defined by game designers, players may choose to adopt their own goals. Researchers interested in using evidence contained in game log files to assess constructs such as persistence have to be careful not to identify a player as lacking persistence when in fact they were very persistently pursuing a different goal. As learning environments become more open-ended and researchers attempt to make inferences about player characteristics from these free-flowing activities, it becomes important to be able to accurately identify players' goals.

DiCerbo & Kidwai (2013) built a detector to classify players based on the pursuit of completion of quests in a game. Using a Classification and Regression Tree methodology, they were able to identify features and rules by which to classify players as serious quest completers with reasonable accuracy compared to human raters. However, completing quests is the major goal intended by game designers, so identifying actions related to this goal is relatively straightforward. This paper takes the group who was not serious about quest completion and attempts to uncover goals for this group based on their game actions.

Method

Participants

Log files from 239 Poptropica players who had been previously judged as not serious about completing quests in the game were analyzed in this study. Players in Poptropica are anonymous, although they do enter gender and grade information. The players in this sample were 48% male and 52% female and ranged in age from 6-14, with a mean age of 10.4 years.

Materials

Poptropica® is a virtual world in which players explore "islands" with various themes and overarching quests that players can choose to pursue. Players choose which islands to visit and navigate the world by walking, jumping and flipping. The quests generally involve completion of 25 or more steps (for example, collecting and using assets) that are usually completed in a particular order. For example, in Vampire's Curse Island (see Figure 1), the player must rescue her friend Katya who has been kidnapped by Count Bram. In order to do this, she must navigate to the Count's castle (eluding wolves on the way), discover the potion for defeating the vampire, identify and mix the potion ingredients, hit the vampire with a potion-tipped arrow from a crossbow, and then find Katya. Apart from the quests, players can talk to other players in highly scripted chats (players can only select from a pre-determined set of statements to make in the chat sessions), play arcade-style games head-to-head, and spend time customizing their avatar with costumes.

Figure 1. Screenshot of Poptropica's Vampire's Curse Island

Like with most online gaming environments, the Poptropica® gaming engine captures time-stamped event data for each player. Events include, for example, the completion of quest-related steps, modifying avatars, collecting objects, and entering new locations. On an average day actions of over 350,000 Poptropica® players generate 80 million event lines. Players can play Poptropica® anonymously and for free or, with purchase of a membership that grants access to premium features such as early access to new quests.

Goal Detection

Establishing Goals. Prior to building a machine detector of goals, it was necessary to establish a standard from which the computer could learn and verify rules. This standard was created by human ratings of log files. Review of a sample of 30 files not part of this analysis revealed that players not completing the quests appeared to fall into two types: explorers and avatar customizers. The explorers group is similar a group suggested by Bartle (1996). Similarly, there has been a significant amount of research on the customization of avatars across virtual worlds (Ducheneaut, Wen, Yee, & Wadley, 2009). These also corresponded to types hypothesized by Poptropica game designers. The ratings were made holistically by examining the full log of each player's actions.

Indentifying features. In order to create a set of *features*—elements of play recorded in the log file—that an automated detector could use to classify a log as representing a serious or not serious attempt to reach the goal of quest completion, it was necessary to define which elements of the log files were indicative of exploration and avatar customization. Poptropica game designers were asked to reflect on cues that might indicate the two types. In this way the following features were identified: (1) total number of scenes (different locations in the game) visited, (2) blimp (way to travel to other parts of the game) opened, (3) return to game map, (4) number of quest events completed (hypothesized to be negatively related to exploration), (5) islands finished (hypothesized to be negatively related to exploration), (6) open costuming tool, and (7) viewing score (points can be used to purchase costumes and hypothesized to be related to customizing the avatar).

Detector Development. Researchers employed a Classification and Regression Tree (CART; Breiman, Friedman, Stone, & Olshen, 1984) methodology to create the detector. CART techniques are particularly useful for developing reliable decision rules to classify individuals or performances into meaningful categories. These techniques require a categorical outcome variable, a set of potential predictor variables, a set of "learning" data by which to establish rules for classification, and a set of "test" data by which to validate those rules. The result of the analysis is a decision tree, or graphical representation of the series of rules for classifying cases. The nodes of the tree denote features, the branches between the nodes give the rules for the values of that feature to be used for classification, and the end nodes of each branch give the final classification of a case. So, one can follow down a series of branches and see if the value of one feature is less than x, and the value of a second feature is more than y, then a person would be classified in group G.

The process of the creation of decision trees begins with the attempt to create classification rules until the data has been categorized as close to perfectly as possible, however, this can result in overfit to the training data. The software then tries to "relax" these rules, in a process called "pruning" to balance accuracy and flexibility to new data. There are a variety of algorithms that can be used to do this pruning in an attempt to find the easiest, most interpretable tree. This research employed the J48 algorithm (Quinlan, 1993), which uses two separate pruning methods to seek solutions, and is a commonly used classification method (Witten & Frank, 2005).

The results of the analyses were evaluated using (1) precision, (2) recall, (3) Cohen's kappa, and (4) A'. Precision is defined as the percent of classified instances that are correct while recall is the percent of total possible instances that are identified (Davis & Goadrich, 2006). Cohen’s kappa provides a measure of agreement between the detector and the raters that is adjusted for chance (it is the same calculation as the one done between two human raters) (Cohen, 1960). A’ is the probability that, given a randomly selected clip, the detector will correctly classify it. It is equivalent to the area under the ROC curve in signal detection theory (Hanley & McNeil, 1982). An A’ of .5 indicates the detector performs as well as chance while an A’ of 1.0 indicates it classifies correctly 100% of the time.

Results

The final decision tree is displayed in Figure 2. Each branch provides classification rules and an ultimate classification decision at the end. The red boxes end paths that indicate avatar customizers while the green boxes end paths that indicate explorers. The numbers in parentheses indicate the number of customizers classified by this rule followed by the number of explorers. So, for example, following the left-most path, we find that people who opened the blimp fewer than 2.5 times should be classified as avatar customizers. The numbers tell us that following this rule would correctly classify 138 players and would not misclassify any. Following other branches reveals different rules, all leading to classifications of customizers or explorers.

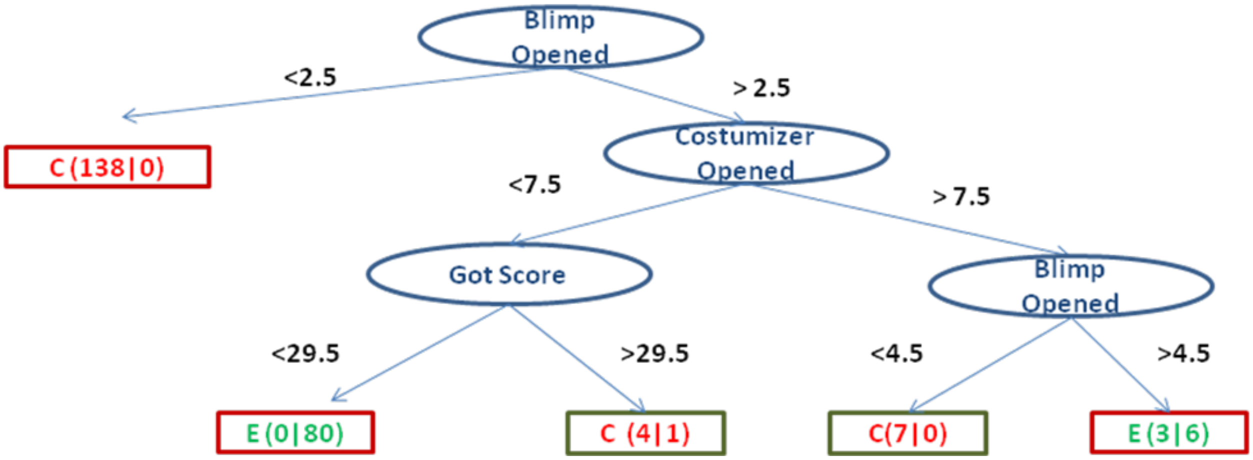

Figure 2. Final Decision Tree

Note: C = Customize avatar, E = Explorer, (Number of customizers|Number of Explorers)

Cross-validation was completed by having half the sample serve as the training sample and the other half as the test sample, and then switching the halves. The detector achieved good performance under cross-validation. A human rater identified 152 of the 239 logs as indicating avatar customization as a goal. The detector identified 150 logs as customization, out of which 149 agreed with the human raters (see Table 1). This resulted in a precision score of .99 and a recall score of .98. The Kappa value was .96, indicating that the accuracy of the detector was 96% better than chance. The A’ was .97, indicating that the detector could correctly classify whether a clip contained serious goal-directed behavior 97% of the time.

Table 1. Correctness of Detector Classification

		Detector	
Human		Customize	Explorer
	Customize	149	3
	Explorer	1	86

Discussion

The purpose of this paper was to describe the creation of a detector of player goals in an online game. The goal a player is pursuing in a game or other open-ended online environment can help provide context and important interpretation of player actions within the system. The results here suggest that an automated detector can be created that can reliably identify whether a participant is engaged in exploration of the environment or customization (in this case of their avatar). The methodology discussed in this paper opens up the possibility of gaming engines detecting and prompting players to adjust their approach (for example, becoming more goal directed) in real time.

Examination of the final decision tree and corresponding analysis reveals that players classified as explorers open the blimp frequently and check their score less frequently. The biggest classifier of customizers was actually the lack of blimp opening, but also those who open the costume tool, and check their score more often. Hypothesized features of number of scenes visited, game map visited, and quest events completed were not helpful in categorization. Given that much of the evidence for customizing is the lack of exploration, future analysis should ensure that there are not other goals these players may be pursuing or other actions that separate this group. However, the current categorization was highly aligned to human judgment.

This type of analysis of player goals may be particularly useful prior to attempts to use log file evidence to measure other constructs. Game-based assessments in particular are likely to result in players who are pursuing a variety of goals. Exploration and customization are elements of games cited as reasons for the use of games in learning and assessment, but they also make assessment more difficult. For example, in work on SimCityEDU, Mislevy et al (2014) attempt to gain evidence of systems thinking from patterns of green energy power placement and pollution removal. However, some players ignore this and begin exploring the effects of bulldozing large portions of the city. The researchers need to pull out this sample because inferences about their understanding of the system may not be valid. Algorithms such as the one discussed in this paper allow for efficient categorization of thousands of players and millions of actions that would not be humanly feasible. It should be clear that scoring game-based assessments involves more than judgments of correct and incorrect. The identification of player goals can help us understand player actions in games and extend our ability to make inferences about player characteristics.

References

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., and Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems,* 402-411.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD Research, 1*, 19.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. London: Chapman & Hall.

Code, J., Clarke-Midura, J., Zap, N. & Dede, C. (2012). Virtual performance assessment for serious games and virtual worlds. In *Interactivity in E-Learning: Cases and Frameworks,* H. Wang, Ed. New York, NY: IGI Publishing.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37-46.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233-240.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology and Society, 17*(1), 17-28. Retrieved from: http://www.ifets.info/journals/17_1/3.pdf

DiCerbo, K. E. & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.) *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273-306). Charlotte, North Carolina: Information Age Publishing.

DiCerbo, K. E. & Kidwai, K. (2013). Detecting Player Goals from Game Log Files. Poster presented at the Sixth International Conference on Educational Data Mining, Memphis, TN.

Hanley, J. & McNeil B. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*, 29-36.

Mislevy, R. J., Oranje, A., Bauer, M. I., vonDavier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K. & John, M. (2014). Psychometric considerations in game-based assessment. [white paper] Retrieved from Institute of Play website: http://www.instituteofplay.org/work/projects/glasslab-research/

Quinlan, R. (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

Rodrigo, M. M. T., Baker, R.S.j.D., McLaren, B.M., Jayme, A. & Dy, T.T. (2012). Development of a workbench to address the educational data mining bottleneck. In: K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper, J. (Eds.) *Proceedings of the 5th International Conference on Educational Data Mining* (EDM 2012), 152-155.

Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2011). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED-2011), Auckland, New Zealand.

Shute, V. (2011). Stealth assessment in computer-based cames to support learning. In *Computer Games and Instruction*, S. Tobias & D. Fletcher, Eds. Information Age Publishers, Charlotte, NC, 503-524.

Shute, V. J. & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In *Design Research on Learning and Thinking in Educational Settings: Enhancing Intellectual*

Growth and Functioning, (pp. 359-387). D. Y. Dai, Ed. New York, NY: Routledge Books.

Ventura, M., & Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*, 2568-2572.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.

Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012). WTF? Detecting students who are conducting inquiry without thinking fastidiously. Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UM