Analyzing Log File Data to Understand Players

Kristen DiCerbo, Pearson, kristen.dicerbo@pearson.com

Abstract: There is great interest in using games to assess players' knowledge, skills, and abilities. However, while games are capable of collecting information about micro-processes of players as they move through the environment, procedures for translating this information into inferences about player proficiencies are not well known. This workshop introduced participants to specific techniques for uncovering patterns in log file data. Techniques were introduced to uncover player groups, action clusters, and patterns of action that illuminate differences between players at different levels of constructs of interest.

Introduction

Digital environments allow us to capture large quantities of information about what students do as they interact with software and each other, seamlessly recorded as they go about their daily activity. These interactions can produce an "ocean" of data, which, if used correctly, can give us a completely different view of how students progress in acquiring knowledge, skills, and attributes (DiCerbo & Behrens, 2012). However, the potential of using this data to understand what students know and can do can be met only if methods for investigating stream or trace data can be developed in psychometrically and computationally feasible ways. Traditional psychometric models have commonly been focused on point-in-time models which overlook variation in activity over time (especially at the micro level). New interactive digital experiences such as on-line learning environments and games, however, elevate both the availability and importance of understanding student temporal micro-patterns which often reflect variation in strategy or evolving psychological states. While the richness of the data holds promise for making important inferences, standard methods for scoring and analysis do not exist. In sum, common practice has not kept up with changes in common data represented in detailed log files.

Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003) helps us link the activities from a game, to evidence gathered from those activities, back to the inferences we want to make about players' knowledge, skills, and abilities. ECD provides the language that lets us articulate the elements needed to gather evidence from game play. First, we need to identify the features of game play that we hypothesize will provide that evidence. Then, we need to determine scoring rules for this evidence. This can be correct/incorrect, present/not present, or a record of the number of seconds taken to complete a task, for example. These pieces of evidence then need to be combined to tell us about the latent, unobservable, constructs we are interested in. There are a variety of statistical methods (e.g., Item Response Theory, Bayesian Networks, Diagnostic Classification Models, and factor analysis) to accomplish this aggregation, depending on the types of data and desired inference.

While ECD helps us define the elements need to make claims about players, in many cases, we have rich log data but weak theory to guide rule formulation. Often, when we design new learning environments, we have some hypothesized models about how actions observed in the game relate to proficiencies, but they are largely a result of expert opinion and best guesses. Without information about what learners actually do in an environment, it is difficult to be confident that we have fully captured the elements of performance that are related to constructs of interest. These conditions suggest that exploratory analyses are needed to uncover these relationships.

Exploratory Data Analysis (EDA) is a conceptual framework aimed at providing insight into data as it is presented to the working researcher (regardless of its origin), and to encourage understanding probabilistic and non-probabilistic models in a way that guards against erroneous conclusions (Behrens, DiCerbo, Levy, & Yel, 2012). It serves to identify patterns and suggest plausible hypotheses to explain them. Using a variety of tools, it encourages the exploration of the patterns in data and the potential explanations for those patterns. Then, the most promising hypotheses can be extracted for further testing using more confirmatory methods.

The purpose of this workshop was to provide examples of projects that had used various methods to make inferences from log file data and introduce participants to tools through that they can use to begin to uncover patterns in log file data.

Review of Existing Projects

The first portion of the session provided examples using a variety of techniques summarized in Table 1.

Aspire - Efficiency	Poptropica – Persistence	SimCity – Information Use
Natural Language Processing – stemming and tagging	Univariate and Bivariate exploratory data analysis	Univariate and Bivariate exploratory data analysis
Identification of Indicators	Classification and Regression Tree	Clustering
Network Analysis	Confirmatory Factor Analysis	Network Analysis
Principal Components Analysis		

Table 1. Game Projects and Constructs with Analysis Techniques

The key take-away from this review was that projects require a range of statistical techniques. The more tools available to the analyst, the greater the flexibility in questions that can be explored and answered.

Getting to Know Data

Following the review of projects, participants were provided with a sample data set and code for analysis. These materials can be obtained from the author. Prior to performing any analyses, data must be cleaned and restructured from its raw form into a form that is suitable for analysis. This often involves selection of cases, examination of outliers, and aggregation of various forms. There are numerous tools available for such work, but some basic Microsoft Excel formulas were provided that accomplish many of these tasks.

The most basic element to examine in data is the distribution of variables. What is the distribution of the number of events that players' engage in during a session? What is the distribution of places visited? Simple histograms are a good place to start with these questions. One thing we usually find is a distribution with some very large outliers, particularly in data related to duration or time spent on activities. In most systems, if someone moves away, engages in another task for a few hours, and comes back, all that is recorded is that there is a 3 hours difference in the times. It is unlikely that someone actually spent three hours on this one event. However, we don't know this from the logs. As a result, we need to select what we think is a maximum reasonable time to complete a task and remove the outliers for most of our analyses regarding time. Participants will examine the distributions, make decisions about the inclusion of outliers, and create new datasets based on these decisions

Following the basic univariate analyses, the most common questions to explore are the relationships between variables. Scatterplot matrices help examine bivariate relationships across a set of variables. These visualizations are essential in combination with numerical correlation analysis for pattern detection as they give insight into the shape (linear, curvilinear, etc.) of the distributions and analysis of relationships between categorical variables.

Finding Groups in Data

One way to look for patterns in data is to examine whether there are groups of people who have similar play patterns or groups of actions that tend to occur together. This is a task for cluster analysis. While the use of clustering of quantitative measures has been well established since the 1930s (Behrens & Smith, 1996), applications of these techniques to categorical log data have only more recently been developed in the Statistical Natural Language Processing (SNLP) literature (Manning & Schuetze, 1999), text mining (Feldman & Sanger, 2007) document retrieval literature, and even the and network security intrusion detection literature (Marchette, 2001). Participants were provided with R code for running hierarchical cluster analysis, including the creation of dendrograms (see Figure 2). However, caution was urged because even the few lines of code required contained a number of assumptions about the algorithms used to compute the clusters. Users should familiarize themselves with the possible choices and implications for the options available.

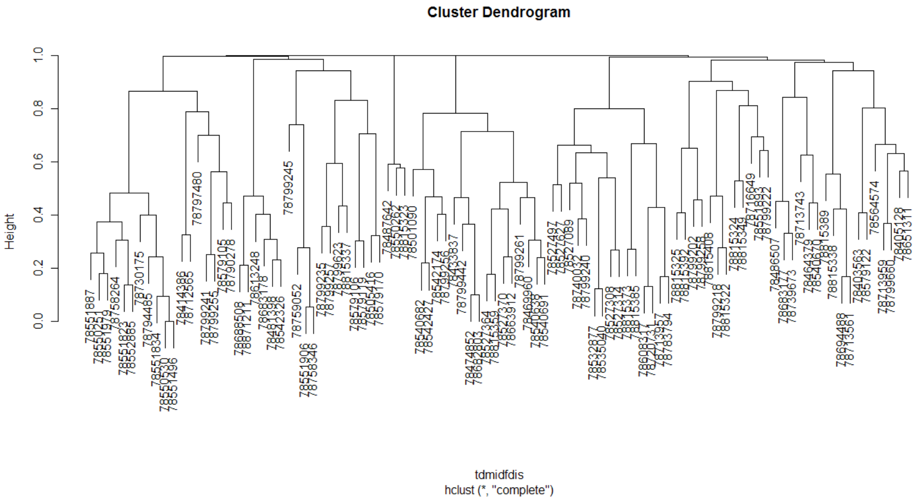

Figure 2: Dendrogram from Hierarchical Cluster Analysis of Users

Analysis of the variables by which players are clustered can reveal the types of measures which may be helpful in differentiating among players at different levels on constructs of interest. For example, we might find clusters of players grouped by time spent on difficult tasks and number of tries after failure that indicate persistence. Similarly, we might find groups based on number of times repeating a challenge and repetition toward optimal solutions that we would hypothesize indicate perfectionism.

Analyzing Sequences of Action

We are often interested in the sequences of actions that players take within a game and there are a number of ways we can analyze these sequences. First, the relatively simple process of creating n-grams, or groups of 2, 3, or n actions to examine the frequencies of the most common sequences of actions will be introduced and undertaken by the participants. Analysis of differences in these frequencies, particularly among groups known to be different on a characteristic of interest, can lead to hypotheses about differences in game play process.

Second, participants will be introduced to the use of sociograms from the tradition of network analysis. The practice of creating sociograms has been used, for example, by teachers in classrooms (Fueyo, & Koorland, 1997), to describe networks of drug users (Pivnick, 1996), and by developmental psychologists (Rodkin, Farmer, Pearl, & Van Acker, 2000). In typical applications, the elements are individuals; people are represented as nodes in the sociogram and links between them as edges. The key to using these techniques in log file analysis is to extend these concepts to thinking of actions as “neighbors” in the network. Each “move” in a log file is more or less connected to other moves, depending on when and how often it is made, as seen in the text mining literature (Feldman & Sanger, 2007). Figure 3 presents sociograms from two players who received the same score in a game. Examination reveals that the paths they took were clearly different. These exploratory analyses can yield important information about how processes of game play can differ even when the final result is similar. In this example, information about how often players switched from one location to another and back, along with their number of total moves, became the basis for a measure of efficiency of solutions.

One of the interesting aspects of sociograms is that social network analysts have developed a series of quantitative measures to describe the relationships seen in the visualization. These include measures of how connected the network is (density), how central an individual is to the network (centrality), and numbers of connections through an individual (degree). These standard metrics of social network analysis may be “translated” to make inferences about the elements of a log file.

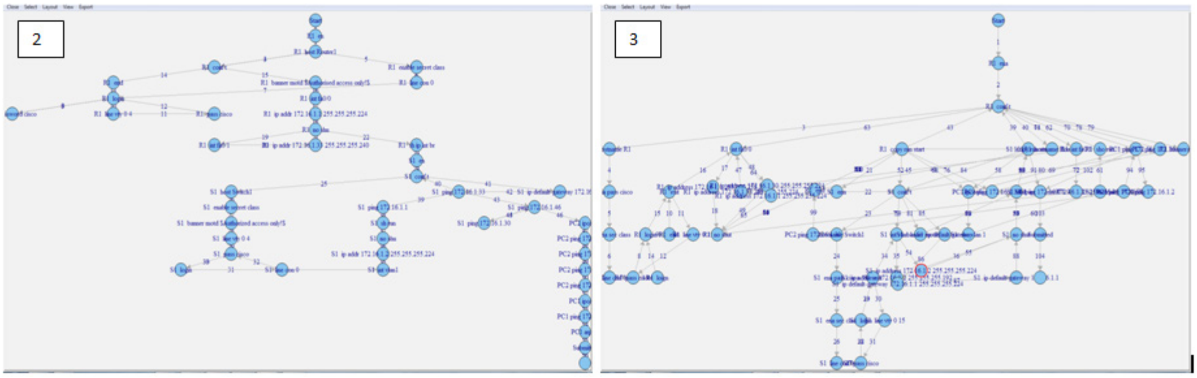

Figure 3: Two sociograms of actions from players who earned the same score

Conclusion

Making inferences from players based on their in-game actions requires the use of a variety of statistical tools and methods, often used in combination. While a brief workshop does not provide sufficient time to introduce, much less master them, the purpose here was to introduce various techniques and provide examples of their application. The work generally requires a frame of exploratory data analysis as hypotheses are rarely developed enough to be explicitly tested in early phases. Tukey (1969) used the analogy of detective work to describe the iterative process of generating hypotheses and looking for fit between facts and the tentative theory or theories, modifying theory and generating new hypotheses. It is important that an emphasis on statistical tools not overshadow the larger meaning-making in this endeavor.

References

Behrens. J. T., DiCerbo, K. E., Yel, N. & Levy, R. (in press). Exploratory data analysis. In W. F. Velicer (Ed.) *Handbook of Psychology: Research Methods in Psychology*. New York: Wiley.

Behrens, J. T., & Smith, M. L. (1996). Data and data analysis. In D. C. Berliner, and R. C. Calfee (Eds.), *Handbook of Educational Psychology*, (pp. 949-989). New York: Macmillan.

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.

Feldman, R. & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Fueyo, V. & Koorland, M. A. (1997). Teacher as researcher: A synonym for professionalism. *Journal of Teacher Education, 48*, 336-344.

Manning, C. D. & Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marchette, D. J. (2001). *Computer intrusion detection and network monitoring: A statistical viewpoint*. New York: Springer-Verlag.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67.

Pivnick, A. (1996). Kinchart-sociograms as a method for describing the social networks of drug-using women. In E. Rahdert (Ed.) Treatment for drug-exposed women and children: Advances in research methodology (pp. 163-182). Washington, DC: National Institutes of Health.

Rodkin P. C., Farmer T. W., Pearl R., & Van Acker, R. (2000). Heterogeneity of popular boys: Antisocial and prosocial configurations. *Developmental Psychology, 36*, 14–24.