Experience Analytics: Developing a Scalable, Implicit and Rich Measure of User Experience

Johanne Christensen  
North Carolina State University  
jtchrist@ncsu.edu

Ben Watson  
North Carolina State University  
bwatson@ncsu.edu

Abstract
New measures of user experience must be defined that can combine the scalability and unobtrusiveness of activity traces with the richness of more traditional measures. Machine learning can be used to predict established UX measures from such activity traces. We advocate research into the type of activity traces needed as input for such measures, the machine learning technology needed, and the user experience components and measures to be predicted.

Author Keywords
human computer interaction; predictive models; supervised learning; web design; user centered design

ACM Classification Keywords
H.1.2 [Models and Principals]: User/Machine Systems

The Need for Improved UX Measures
Companies such as Facebook and Google have long understood the value of the data they collect, and have pioneered its use in measuring experience [Müller and Sedley 2014][Rodden et al. 2010]. By improving experience, they garner more user attention, and can improve ad sales. Politicians have also pioneered such analytics [Anderson and Horvath 2016]. By measuring reaction to their messages, they hope to target ads more...
effectively, and gain votes.

Experiential data collected with such scale and efficiency may have promise for many of society's most pressing behavioral problems: health conditions such as obesity and diabetes; security problems like social engineering; and changes in sustainability habits, like using less power at home. But understanding what users are attending to may not be enough to realize this promise. Instead, we need to understand what such data implies about user emotion, intent and behavior.

Traditionally, researchers gain this deeper experiential understanding using surveys and interviews. Unfortunately, the rapid iteration of modern application development leaves little time for using such measures at meaningful scale. In addition, because they alter the experience being measured, surveys and interviews may suffer from the observer effect [Landsberger 1958][Cukor-Avila 2000], calling their validity into question.

Building a New Measure of User Experience
Could we combine the scalability and unobtrusiveness of web analytics with the rich meaning of traditional measures? That is, could we predict nuanced emotion, intent and behavior from large-scale, implicitly collected behavior traces? We believe so.

We propose experience analytics: a scalable and implicit technology enabling reliable and detailed prediction of rich user experience. Web analytics has at least tacitly promised such prediction for years, however, recent work by [O'Brien and Lebow 2013] suggests that simple statistics drawn from such data are not good predictors of the user's actual experience. By applying machine learning to behavioral data to predict survey results, experience analytics can connect simple statistics into more meaningful patterns of behavior, and draw inferences about the underlying experience that created these behavioral patterns. We have reason to be optimistic: [Fox et al. 2005] have found success predicting user satisfaction during web search using such modeling techniques, and [Epp et al. 2011] predicted emotional state from keystrokes.

What types of behavioral data will be useful as input?
Useful data for experience analytics should be:

- **Easily scalable.** If large data sets are required, then data collection and analysis will only be practical if it can be easily automated.
- **Implicitly collected.** Whenever possible, experience analytics data should be collected unobtrusively, to avoid the possibility of affecting behavior.
- **Rich with meaning.** Data should allow deep inferences about user experience such as liking, engagement, curiosity and confusion — and connect this experience to particular user behaviors and system components.

Surveys are widely utilized in UX, and perhaps most familiar to users as popups that appear when visiting a website. Such surveys continue to be improved and refined [O'Brien and Cairns 2015]. [Müller and Sedley 2014] are exploring ways to deploy at a surveys at large scale. UX researchers like [Rodden et al. 2010] are testing methods of combining them with other metrics. With any of several maturing toolsets, surveys can produce data in a scalable fashion. However, surveys are still intrusive, which can result in poor participation and unbalanced sampling. Moreover, they rely on a user's self-reflection, which is itself often unreliable.
Experience can be also be inferred from direct measurement of the user, usually performed in a lab. However, technical and privacy constraints mean that most of these technologies will never be scalable, and many are quite intrusive. Moreover, most of these measures are relatively new, and do not have a well established mapping from data to experience.

Behavior traces meet all three of our criteria. Tools like Google Analytics make it easy to collect traces in a scalable fashion. As with web cookies, users are at most peripherally aware that their behavior is being traced (they are certainly not interrupted with modal dialogs). Commonly tracked behaviors include time spent on page, clicks, scrolling, etc. Cursor tracking is a less commonly traced behavior, but may add richness to the data set and has been shown to be useful in determining a user's goals on the website [Guo and Agichtein 2010][Warnock and Lalmas 2015].

This begs two further questions. Firstly, which type of behavioral data will be most predictive? We view this as an open question that can only be answered by the research we advocate. As we find answers, we might record and analyze only those components that we find relevant. Secondly, how should this data be aggregated? We might group data by time (one session, a week, or longer), user (one user, a certain demographic, or all users), or sections of the website or app (a single page, a section, or the entire site). Ultimately, the aggregation depends on the amount of data available, the questions being posed in analysis, as well as the structure of the website or app being analyzed. Therefore, we will likely want to include tuning options in order to produce the best possible model for each scenario.

Possible technologies for making experiential predictions [O’Brien and Lebow 2013] speculated that simple aggregations of standard web analytics lack the contextual information necessary to differentiate between experiences. We propose exploring interactions of behavioral traces in their system and task context. This approach builds a richer understanding of the data, and will enable us to make better predictions of the user's underlying experience.

Which type of machine learning will prove most predictive remains to be seen. However, research on web search is telling, and suggests that classification algorithms will be a useful starting point. [Feild et al. 2010] successfully used logistic regression to model user frustration from a combination of query log data and physical sensor data. [Fox et al. 2005] used Bayesian networks and decision trees with behavior traces as input, and reliably predicted the data collected using surveys measuring user interest and satisfaction. Decision trees have been used to evaluate behavioral traces [Velayathan and Yamada 2006], by comparing behavior to a previously built profile of past behavior to determine the user's interest in a web page.

Which components of experience should we predict? If we wish to predict user experience, how might we reliably measure that experience in a manner that works with ML? There are several important possibilities, each a well-known component of user experience, and each quite complex:

- **Satisfaction**: perhaps the most well established component of user experience. There are several reliable surveys for measuring satisfaction, including the Questionnaire for User Interface Satisfaction (QUIS) [Chin et al. 1988]. A recent alternative uses image processing to detect satisfaction from facial expression [Kasiran and Yahya 2007].

- **Engagement**: engagement has found currency in research as advertising has moved to web and mobile platforms. One standard measure that has emerged
is the User Engagement Survey (UES) [O’Brien and Toms 2010]. Less explicit measures have been proposed, but are largely untested.

- **Emotion**: emotions are complex, their relationship to experience equally complicated, and they have only recently received deeper attention in UX research [Hassenzahl et al. 2010][Dillard and Seo 2012]. There are several well established survey measures, such as the Self Assessment Manikin (SAM) [Bradley and Lang 1994]. Recently, several behavioral and physiological measures have been proposed, including facial expression, GSR and EEG.

- **Intent**: in the very mature study of persuasion [O’Keefe 2012], intent is a proxy for actual behavior, which is much more difficult to capture reliably. Typically this is measured with surveys customized to the persuasive objective.

- **Behavior**: studying behavior has either required following users into the field, or asking them to answer retrospectively and unreliably [Rhodes and D.R 2012]. However, on web and mobile platforms, behavior can either be measured directly (e.g. clicking "buy"), or with sensors (e.g. GPS to log method of transport).

As we build our models of experience, we should keep in mind that because of their implicit nature, behavioral traces may not be the best basis for predicting intrusive surveys. This would not necessarily mean that traces are bad predictors of experience: they are related, but not the same.

**Validating Our New UX Measure**

To test and validate our new measure of user experience, we will compare it to more traditional survey measures, as the type of application and interface vary. To vary the type of application, we are gathering data from different types of websites, public kiosks, and vehicle dashboards. To vary interface, we are creating pretested interfaces that deliver good and bad experiences.

Initially, our experiments will focus on satisfaction, as measured in QUIS. As our work progresses, we plan develop new trace-based measures that predict more nuanced components of UX.

We will evaluate several types of validity while testing our measure. To examine construct validity, we compare factor analyses of our new measure and QUIS. To examine construct validity, we will rely on the well-established user satisfaction literature (e.g. [Chin et al. 1988]), striving to ensure that our new measure has components related to all the known latent factors of user satisfaction. Finally, to verify criterion validity, we will compare our new measure’s response to good and bad interfaces with that of QUIS. The two should be correlated, however, because there is currently no validated implicit measure to compare to our new implicit measure, we do not expect this correlation to be particularly high.

**Conclusion**

Better utilization of data describing user behavior is a necessary next step in improving UX. We believe that such scalable, implicit and nuanced UX measurement technology will not only improve the corporate bottom line, but help society solve behavioral problems such as health and sustainability by giving the public access to many of the same user experience measurement techniques previously available only to corporate conglomerates.
REFERENCES


