Measuring the User Experience on a Large Scale: User-Centered Metrics for Web Applications

Kerry Rodden, Hilary Hutchinson, and Xin Fu
Google
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
{krodden, hhutchinson, xfu}@google.com

ABSTRACT
More and more products and services are being deployed on the web, and this presents new challenges and opportunities for measurement of user experience on a large scale. There is a strong need for user-centered metrics for web applications, which can be used to measure progress towards key goals, and drive product decisions. In this note, we describe the HEART framework for user-centered metrics, as well as a process for mapping product goals to metrics. We include practical examples of how HEART metrics have helped product teams make decisions that are both data-driven and user-centered. The framework and process have generalized to enough of our company's own products that we are confident that teams in other organizations will be able to reuse or adapt them. We also hope to encourage more research into metrics based on large-scale behavioral data.

Author Keywords
Metrics, web analytics, web applications, log analysis.

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces—benchmarking, evaluation/methodology.

General Terms
Experimentation, Human Factors, Measurement.

INTRODUCTION

Advances in web technology have enabled more applications and services to become web-based and increasingly interactive. It is now possible for users to do a wide range of common tasks "in the cloud", including those that were previously restricted to native client applications (e.g. word processing, editing photos). For user experience professionals, one of the key implications of this shift is the ability to use web server log data to track product usage on a large scale. With additional instrumentation, it is also possible to run controlled experiments (A/B tests) that compare interface alternatives. But on what criteria should they be compared, from a user-centered perspective? How should we scale up the familiar metrics of user experience, and what new opportunities exist?

In the CHI community, there is already an established practice of measuring attitudinal data (such as satisfaction) on both a small scale (in the lab) and a large scale (via surveys). However, in terms of behavioral data, the established measurements are mostly small-scale, and gathered with stopwatches and checklists as part of lab experiments, e.g. effectiveness (task completion rate, error rate) and efficiency (time-on-task) [13].

A key missing piece in CHI research is user experience metrics based on large-scale behavioral data. The web analytics community has been working to shift the focus from simple page hit counts to key performance indicators. However, the typical motivations in that community are still largely business-centered rather than user-centered. Web analytics packages provide off-the-shelf metrics solutions that may be too generic to address user experience questions, or too specific to the e-commerce context to be useful for the wide range of applications and interactions that are possible on the web.

We have created a framework and process for defining large-scale user-centered metrics, both attitudinal and behavioral. We generalized these from our experiences of working at a large company whose products cover a wide range of categories (both consumer-oriented and business-oriented), are almost all web-based, and have millions of users each. We have found that the framework and process have been applicable to, and useful for, enough of our company's own products that we are confident that teams in other organizations will be able to reuse or adapt them successfully. We also hope to encourage more research into metrics based on large-scale behavioral data, in particular.

RELATED WORK

Many tools have become available in recent years to help with the tracking and analysis of metrics for web sites and applications. Commercial and freely available analytics

packages [5,11] provide off the shelf solutions. Custom analysis of large-scale log data is made easier via modern distributed systems [4,8] and specialized programming languages [e.g. 12]. Web usage mining techniques can be used to segment visitors to a site according to their behavior [3]. Multiple vendors support rapid deployment and analysis of user surveys, and some also provide software for large-scale remote usability or benchmarking tests [e.g. 14]. A large body of work exists on the proper design and analysis of controlled A/B tests [e.g. 10] where two similar populations of users are given different user interfaces, and their responses can be rigorously measured and compared.

Despite this progress, it can still be challenging to use these tools effectively. Standard web analytics metrics may be too generic to apply to a particular product goal or research question. The sheer amount of data available can be overwhelming, and it is necessary to scope out exactly what to look for, and what actions will be taken as a result. Several experts suggest a best practice of focusing on a small number of key business or user goals, and using metrics to help track progress towards them [2, 9, 10]. We share this philosophy, but have found that this is often easier said than done. Product teams have not always agreed on or clearly articulated their goals, which makes defining related metrics difficult.

It is clear that metrics should not stand alone. They should be triangulated with findings from other sources, such as usability studies and field studies [6,9], which leads to better decision-making [15]. Also, they are primarily useful for evaluation of launched products, and are not a substitute for early or formative user research. We sought to create a framework that would combine large-scale attitudinal and behavioral data, and complement, not replace, existing user experience research methods in use at our company.

PULSE METRICS

The most commonly used large-scale metrics are focused on business or technical aspects of a product, and they (or similar variations) are widely used by many organizations to track overall product health. We call these PULSE metrics: **P**age views, **U**ptime, **L**atency, **S**even-day active users (i.e. the number of unique users who used the product at least once in the last week), and **E**arnings.

These metrics are all extremely important, and are related to user experience – for example, a product that has a lot of outages (low uptime) or is very slow (high latency) is unlikely to attract users. An e-commerce site whose purchasing flow has too many steps is likely to earn less money. A product with an excellent user experience is more likely to see increases in page views and unique users.

However, these are all either very low-level or indirect metrics of user experience, making them problematic when used to evaluate the impact of user interface changes. They may also have ambiguous interpretation – for example, a rise in page views for a particular feature may occur because the feature is genuinely popular, or because a confusing interface leads users to get lost in it, clicking around to figure out how to escape. A change that brings in more revenue in the short term may result in a poorer user experience that drives away users in the longer term.

A count of unique users over a given time period, such as seven-day active users, is commonly used as a metric of user experience. It measures the overall volume of the user base, but gives no insight into the users' level of commitment to the product, such as how frequently each of them visited during the seven days. It also does not differentiate between new users and returning users. In a worst-case retention scenario of 100% turnover in the user base from week to week, the count of seven-day active users could still increase, in theory.

HEART METRICS

Based on the shortcomings we saw in PULSE, both for measuring user experience quality, and providing actionable data, we created a complementary metrics framework, HEART: **H**appiness, **E**ngagement, **A**doption, **R**etention, and **T**ask success. These are categories, from which teams can then define the specific metrics that they will use to track progress towards goals. The Happiness and Task Success categories are generalized from existing user experience metrics: Happiness incorporates satisfaction, and Task Success incorporates both effectiveness and efficiency. Engagement, Adoption, and Retention are new categories, made possible by large-scale behavioral data.

The framework originated from our experiences of working with teams to create and track user-centered metrics for their products. We started to see patterns in the types of metrics we were using or suggesting, and realized that generalizing these into a framework would make the principles more memorable, and usable by other teams.

It is not always appropriate to employ metrics from every category, but referring to the framework helps to make an explicit decision about including or excluding a particular category. For example, Engagement may not be meaningful in an enterprise context, if users are expected to use the product as part of their work. In this case a team may choose to focus more on Happiness or Task Success. But it may still be meaningful to consider Engagement at a feature level, rather than the overall product level.

Happiness

We use the term "Happiness" to describe metrics that are attitudinal in nature. These relate to subjective aspects of user experience, like satisfaction, visual appeal, likelihood to recommend, and perceived ease of use. With a general, well-designed survey, it is possible to track the same metrics over time to see progress as changes are made.

For example, our site has a personalized homepage, iGoogle. The team tracks a number of metrics via a weekly in-product survey, to understand the impact of changes and

new features. After launching a major redesign, they saw an initial decline in their user satisfaction metric (measured on a 7-point bipolar scale). However, this metric recovered over time, indicating that change aversion was probably the cause, and that once users got used to the new design, they liked it. With this information, the team was able to make a more confident decision to keep the new design.

Engagement

Engagement is the user's level of involvement with a product; in the metrics context, the term is normally used to refer to behavioral proxies such as the frequency, intensity, or depth of interaction over some time period. Examples might include the number of visits per user per week, or the number of photos uploaded per user per day. It is generally more useful to report Engagement metrics as an average per user, rather than as a total count – because an increase in the total could be a result of more users, not more usage.

For example, the Gmail team wanted to understand more about the level of engagement of their users than was possible with the PULSE metric of seven-day active users (which simply counts how many users visited the product at least once within the last week). With the reasoning that engaged users should check their email account regularly, as part of their daily routine, our chosen metric was the percentage of active users who visited the product on five or more days during the last week. We also found that this was strongly predictive of longer-term retention, and therefore could be used as a bellwether for that metric.

Adoption and Retention

Adoption and Retention metrics can be used to provide stronger insight into counts of the number of unique users in a given time period (e.g. seven-day active users), addressing the problem of distinguishing new users from existing users. Adoption metrics track how many new users start using a product during a given time period (for example, the number of accounts created in the last seven days), and Retention metrics track how many of the users from a given time period are still present in some later time period (for example, the percentage of seven-day active users in a given week who are still seven-day active three months later). What counts as "using" a product can vary depending on its nature and goals. In some cases just visiting its site might count. In others, you might want to count a visitor as having adopted a product only if they have successfully completed a key task, like creating an account. Like Engagement, Retention can be measured over different time periods – for some products you might want to look at week-to-week Retention, while for others monthly or 90-day might be more appropriate. Adoption and Retention tend to be especially useful for new products and features, or those undergoing redesigns; for more established products they tend to stabilize over time, except for seasonal changes or external events.

For example, during the stock market meltdown in September 2008, Google Finance had a surge in both page views and seven-day active users. However, these metrics did not indicate whether the surge was driven by new users interested in the crisis, or existing users panic-checking their investments. Without knowing who was making more visits, it was difficult to know if or how to change the site. We looked at Adoption and Retention metrics to separate these user types, and examine the rate at which new users were choosing to continue using the site. The team was able to use this information to better understand the opportunities presented by event-driven traffic spikes.

Task Success

Finally, the "Task Success" category encompasses several traditional behavioral metrics of user experience, such as efficiency (e.g. time to complete a task), effectiveness (e.g. percent of tasks completed), and error rate. One way to measure these on a large scale is via a remote usability or benchmarking study, where users can be assigned specific tasks. With web server log file data, it can be difficult to know which task the user was trying to accomplish, depending on the nature of the site. If an optimal path exists for a particular task (e.g. a multi-step sign-up process) it is possible to measure how closely users follow it [7].

For example, Google Maps used to have two different types of search boxes – a dual box for local search, where users could enter the "what" and "where" aspects separately (e.g. [pizza][nyc]) and a single search box that handled all kinds of searches (including local searches such as [pizza nyc], or [nyc] followed by [pizza]). The team believed that the single-box approach was simplest and most efficient, so, in an A/B test, they tried a version that offered only the single box. They compared error rates in the two versions, finding that users in the single-box condition were able to successfully adapt their search strategies. This assured the team that they could remove the dual box for all users.

GOALS – SIGNALS – METRICS

No matter how user-centered a metric is, it is unlikely to be useful in practice unless it explicitly relates to a goal, and can be used to track progress towards that goal. We developed a simple process that steps teams through articulating the *goals* of a product or feature, then identifying *signals* that indicate success, and finally building specific *metrics* to track on a dashboard.

Goals

The first step is identifying the goals of the product or feature, especially in terms of user experience. What tasks do users need to accomplish? What is the redesign trying to achieve? Use the HEART framework to prompt articulation of goals (e.g. is it more important to attract new users, or to encourage existing users to become more engaged?). Some tips that we have found helpful:

- Different team members may disagree about what the project goals are. This process provides a great

opportunity to collect all the different ideas and work towards consensus (and buy-in for the chosen metrics).

- Goals for the success of a particular *project* or *feature* may be different from those for the *product* as a whole.
- Do not get too distracted at this stage by worrying about whether or how it will be possible to find relevant signals or metrics.

Signals

Next, think about how success or failure in the goals might manifest itself in user behavior or attitudes. What actions would indicate the goal had been met? What feelings or perceptions would correlate with success or failure? At this stage you should consider what your data sources for these signals will be, e.g. for logs-based behavioral signals, are the relevant actions currently being logged, or could they be? How will you gather attitudinal signals – could you deploy a survey on a regular basis? Logs and surveys are the two signal sources we have used most often, but there are other possibilities (e.g. using a panel of judges to provide ratings). Some tips that we have found helpful:

- Choose signals that are sensitive and specific to the goal – they should move only when the user experience is better or worse, not for other, unrelated reasons.
- Sometimes failure is easier to identify than success (e.g abandonment of a task, "undo" events [1], frustration).

Metrics

Finally, think about how these signals can be translated into specific metrics, suitable for tracking over time on a dashboard. Some tips that we have found helpful:

- Raw counts will go up as your user base grows, and need to be normalized; ratios, percentages, or averages per user are often more useful.
- There are many challenges in ensuring accuracy of metrics based on web logs, such as filtering out traffic from automated sources (e.g. crawlers, spammers), and ensuring that all of the important user actions are being logged (which may not happen by default, especially in the case of AJAX or Flash-based applications).
- If it is important to be able to compare your project or product to others, you may need to track additional metrics from the standard set used by those products.

CONCLUSIONS

We have spent several years working on the problem of developing large-scale user-centered product metrics. This has led to our development of the HEART framework and the Goals-Signals-Metrics process, which we have applied to more than 20 different products and projects from a wide variety of areas within Google. We have described several examples in this note of how the resulting metrics have helped product teams make decisions that are both data-driven and user-centered. We have also found that the framework and process are extremely helpful for focusing discussions with teams. They have generalized to enough of our company's own products that we are confident that teams in other organizations will be able to reuse or adapt them successfully. We have fine-tuned both the framework and process over more than a year of use, but the core of each has remained stable, and the framework's categories are comprehensive enough to fit new metrics ideas into. Because large-scale behavioral metrics are relatively new, we hope to see more CHI research on this topic – for example, to establish which metrics in each category give the most accurate reflection of user experience quality.

ACKNOWLEDGMENTS

Thanks to Aaron Sedley, Geoff Davis, and Melanie Kellar for contributing to HEART, and Patrick Larvie for support.

REFERENCES

1. Akers, D. et al. (2009). Undo and Erase Events as Indicators of Usability Problems. *Proc of CHI 2009*, ACM Press, pp. 659-668.
2. Burby, J. & Atchison, S. (2007). Actionable Web Analytics. Indianapolis: Wiley Publishing, Inc.
3. Chi, E. et al. (2002). LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition. *Proc of WebKDD 2002,* ACM Press, pp. 1-15.
4. Dean, J. & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM, 51 (1)*, pp. 107-113.
5. Google Analytics: http://www.google.com/analytics
6. Grimes, C. et al. (2007). Query Logs Alone are not Enough. *Proc of WWW 07 Workshop on Query Log Analysis*: http://querylogs2007.webir.org
7. Gwizdka, J. & Spence, I. (2007). Implicit Measures of Lostness and Success in Web Navigation. *Interacting with Computers* 19(3), pp. 357-369.
8. Hadoop: http://hadoop.apache.org/core
9. Kaushik, A. (2007). Web Analytics: An Hour a Day. Indianapolis: Wiley Publishing, Inc.
10. Kohavi, R. et al. (2007). Practical Guide to Controlled Experiments on the Web. *Proc of KDD 07*, ACM Press, pp. 959-967.
11. Omniture: http://www.omniture.com
12. Pike, R. et al. (2005). Interpreting the Data: Parallel Analysis with Sawzall. *Scientific Programming (13)*, pp. 277-298.
13. Tullis, T. & Albert, W. (2008). *Measuring the User Experience*. Burlington: Morgan Kaufmann.
14. UserZoom: http://www.userzoom.com
15. Weischedel, B. & Huizingh, E. (2006). Website Optimization with Web Metrics: A Case Study. *Proc of ICEC 06*, ACM Press, pp. 463-470.