Pre-reading for Information Analytics Case Study
SI310 Case 9
Fall 2019
Written by Kevyn Collins-Thompson

It's difficult to overstate how critical the ability to analyze, interpret and exploit Web log data has become in today's information economy. This pre-reading provides background showing how Web log data is a rich online resource for understanding and predicting important properties and actions of both customers and online providers, and knowing what kinds of biases and tradeoffs one must think about in using such data.

Basic data fields and statistics. The most basic information in a Web log is a record of the URL for each page that users visit on the site, along with a timestamp recording exactly when the URL visit was started. The most common type of interaction users have with a site is clicking to perform an action (e.g. clicking a button to purchase, clicking a link for more information). Each click results in a URL request, which is recorded as an entry in the log. Typically, customers can have an account or login ID for the site, and a unique identifier associated with their account is also stored with that URL. By examining the trail of URLs a given individual user visits as they progress through a site, the organization can interpret how a user is making use of the site's features and content. There are ways to encode multiple pieces of information into a URL, so that not only can each item in a database have a corresponding URL that's associated with an action on that item, but different variables about the context of the action (e.g. if this click is coming from a special promotion page vs. a regular page) can also be encoded in a URL. As another example, it is also easy, by adding a little Javascript to a site's pages, to record the stream of mouse pointer positions that comes from a user's mouse movements, which are then sent to the server as a sequence of URLs with specially encoded extra data. Important to note is that if a user clicks a link to content that is not on the site in question, the system can record the initial click on the site's page, but any requests a user might make to a different, second site is not typically recorded by the first site (although as you might expect, there are certainly sneaky ways to record this if a company wanted this data). If you read the subsection '*Sources of Web Use-Based Data*' of the Burton and Walther article [2], you'll learn more specifics on what information is gathered from Web log data (in the context of one key task: evaluating a Web site's design).

Customer metrics. Next, given a trail of site actions by a customer, there are many customer metrics that can be derived, the estimating of which in fact may be critical for understanding and predicting a company's 'bottom line' profitability. Most broadly, for example, an online retailer might measure statistics like total sales in dollars from online purchases, visitor interest in certain items as measured by searches, and the 'click through' rate at which customers click on an online advertisement and end up buying the product or service. By monitoring users' progress through an expected list of URLs, a site that allows people to complete certain tasks

might measure the percentage of tasks that were started (e.g. by recording visits to a task start page) vs successfully or unsuccessfully completed (e.g. hitting 'Submit' without errors on the final task web page to open an account). If customers have a unique user ID, then many user-specific metrics can be estimated, such as the customer return rate to the website. Even further, if a site has demographic or financial data about a user, then this can be joined with the unique user ID, so that any of the customer metrics can be faceted (put into different 'buckets') by any user variable to get a picture for how that metric varies across different types of users. For example, a brokerage firm could identify the most popular investment bought by females aged 40-49 living in Toronto, Canada – and target future advertising appropriately. Often, no single metric can capture a complete picture of how users are interacting with the website. That's why a company's evaluation team will typically use a scorecard of 10 to 20 different metrics simultaneously. The exact makeup of the scorecard will reflect the variables that the organization views as needing tradeoffs – for example, an e-commerce company might trade off the value of keeping users engaged during a longer site visit (gathering better information that could be used, e.g. to help with their task, offer more products, show relevant ads) vs. satisfying their needs quickly (in hopes this makes them happy and eager to return again). See the chapter reading by Trites, Bortiz and Pugsley [5] for an explanation of metrics and their data sources for e-commerce companies (it's written with a slight orientation towards a Canadian perspective: but you don't need to cover the boxed examples. The core material is applicable everywhere).

Sources of bias. Bias is often seen as "bad" but there are good biases. For example, we are biased in favor of good work. In some cases, however, bias can cause problems and make work bad. The trick is to spot the biases that work against being good. These can occur in any analytical task, but given the stakes involved in web log analysis, bias is especially important to consider when trying to conduct analyses for a variety of stakeholders, all of whom may have different interests and expectations. In choosing how to do an analysis, and then how to interpret the results of one, many sources of potential bias can influence and corrupt the conclusions that can be drawn from the data.

One of the most important, widespread, and easy to miss sources of bias that can be bad is confirmation bias, which happens when someone has decided ahead of time what conclusions they would like to reach from the data, picking only those results that support their pre-determined conclusion, and ignoring any evidence to the contrary. Michael Walker's blog article [4] discusses confirmation bias in the context of the Data Science Code of Professional Conduct, which gives guidelines that help data science practitioners – including those working with web log data - produce analyses that are informative, credible, and ethical.

A second form of bias that can be bad and is critical to understand is availability bias, which causes us to overestimate how likely memorable or dramatic events are. With web log data, your analysis might have found an occurrence where a user was given exactly the right suggestion for a stock to buy, resulting in huge financial gains. However, such events were also

exceedingly rare, so citing this as evidence that the suggestion algorithm was effective ignores the truth about the overall behavior of the algorithm: that most users are getting ineffective suggestions.

A third critical type of bias that can be bad is selection bias, which occurs when, for example, you choose to analyze a data sample that isn't representative of your users as a whole. For example, you might decide to analyze only site visitors who use one of the top 3 most popular types of browser. However, if these browser types were strongly associated with North America (by visit count), you might then be completely ignoring all users from, say, a specific overseas country (and important future customers) where a different browser is very popular. Sometimes, confirmation bias is enabled by different types of selection bias when gathering evidence.

There other important types of bias that can be bad, such as training bias, that applies to algorithms that attempt to make predictions from data. However, these are beyond the scope of this case. (The project courses deal with such things.)

Online experiments. A final key role of web log data is that it captures the results of online experimentation that companies use to help them understand how features of their web site may affect important customer metrics. For example, Amazon relies on constant online experimentation to understand the effect of changes and improvements. Some changes might be minor, e.g. adding a border a few pixels wide around the picture of a product in a product description. Other changes might be more significant, such as using a new product recommendation algorithm.

Instead of simply deploying new changes without testing, organizations will often perform online experiments with a small fraction of their users to determine the value of possible changes. The most basic kind of online experimentation is A-B testing, so called because users are randomly assigned to one of two distinct groups (A or B), each of which is shown a different version of the site. Often, user group A represents a control condition (say, the existing version of the site) and group B represents a 'treatment' condition (say, a modified version of the site with a better recommendation algorithm). By comparing the customer metric generated for group A against group B, the company can decide whether the new recommendation feature may be worth keeping on the site.

All user interactions with the site that comprise this experiment data are captured as part of the log database, typically by storing an experiment ID with each log entry of a page shown, and a table storing the mapping of each user to the experiment ID. Kohavi and Thomke [1] give an excellent overview of online experimentation along with some compelling real-world examples of what can be learned from this kind of web log analysis. Both authors are highly experienced industry veterans with fascinating stories to tell about the value of Web log data from online experiments.

So much sensitive data is now stored as part of online web site interactions that clear privacy, security, and ethics issues associated with collecting and using web log data arise. There are also real ethical and privacy concerns in designing and using the data from online experiments that must be balanced against the usefulness of the experiment: see the TechCrunch article [3] for a widely-reported example from Facebook. In addition, by joining different sources of data about a user, a highly revealing and detailed picture of an individual's interests, habits, and behavior can be built from web log data. Users can also reveal personally identifiable information directly through use of features such as search queries on a site, so having those queries appear as 'suggested queries' for other users, or even as examples in internal reports, would not be desirable. In another example, systems that attempt to personalize product recommendations based on past purchase data may "leak" information about a user's preferences that they may not want made public or visible.

References

[1] R. Kohavi, S. Thomke. The Surprising Power of Online Experiments. Harvard Business Review. Sept/Oct. 2017. https://hbr.org/2017/09/the-surprising-power-of-online-experiments

[2] Mary C. Burton, Joseph B. Walther. The Value of Web Log Data in Use-Based Design and Testing. J. of Computer-Mediated Communication http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2001.tb00121.x/full

[3] J. Constine. The Morality of A/B Testing. TechCrunch. Retrieved Oct 23, 2017. https://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/

[4] M. Walker. The Deadly Data Science Sin of Confirmation Bias. Data Science Central. Retrieved Oct 23, 2017. https://www.datasciencecentral.com/profiles/blogs/the-deadly-data-science-sin-of-confirmation-bias

[5] G. Trites, J. Bortiz, D. Pugsley. Metrics for Performance Measurement in E-Commerce. Chapter 12 in *E-Business: A Canadian Perspective for a Networked World*. Second edition. Pearson. http://www.pearsoned.ca/highered/divisions/virtual_tours/trites/data/Trites_EBus_Ch12.pdf