WEPIR 2018 Keynote Address
Offline Evaluation of Personalization Using Logged Data
Ben Carterette

Abstract

Evaluation of algorithms for personalization is very important for being able to iteratively develop improved algorithms, but it is not always easy to do. Batch experimentation using a test collection is fast, but has high start-up costs, often requires very strong assumptions about users and their needs in context, and can introduce biases if the data has not been collected very carefully. User studies are slow and have high variance, making them difficult to generalize and certainly not possible to use for iterative development. Online experimentation using A/B tests, pioneered and refined by companies such as Google and Microsoft, can be fast but is limited in other ways, in particular that it is not easy to do without access to a large user base.

In this talk I present work we have done and work in progress on using logged online user data to evaluate personalization offline. I will discuss some of the user simulation work I have done with my students in the context of evaluating system effectiveness over user search sessions (in the context of the TREC Session track), based on training models on logged data for use offline. I will also discuss work on using historical logged data to re-weight search outputs for evaluation, focusing on how to collect that data to arrive at unbiased conclusions. The latter is work I am doing while on sabbatical at Spotify, which provides many motivating examples.

Biographical Sketch

Ben Carterette is an Associate Professor in the Department of Computer and Information Sciences at the University of Delaware, and currently on sabbatical as a Research Scientist at Spotify in New York City. He primarily researches search evaluation, including everything from designing search experiments to building test collections to obtaining relevance judgments to using them in evaluation measures to statistical testing of results. He completed his PhD with James Allan at the University of Massachusetts Amherst on low-cost methods for acquiring relevance judgments for IR evaluation. He has published over 80 papers, won 4 Best Paper Awards, and co-organized two ACM SIGIR-sponsored conferences--WSDM 2014 and ICTIR 2016--in addition to nearly a decade's worth of TREC tracks and several workshops on topics related to new test collections and evaluation. He was also elected SIGIR Treasurer in 2016.