Challenges and Limitations in the Offline and Online Evaluation of Recommender Systems: A Netflix Case Study

Carlos Gomez-Uribe
Netflix, USA
cgomez@netflix.com

ABSTRACT

The typical use case of recommendation systems is suggesting items such as videos, songs or articles to users. Evaluating a recommender system is critical to the process of improving it. In theory the best judges of the quality and effectiveness of a recommender system are the users themselves, e.g., ideal metrics can describe the intensity and frequency of a user's interaction with the system over the long term. In practice, however, despite the wide adoption of consumer science based on online A/B testing for the evaluation and comparison of different recommender systems, user-derived measurements are often noisy, slow, non-repeatable, and sensitive to a myriad of potential confounders. Furthermore, conducting large-scale user experiments for researchers in academia is often impossible. A complementary offline approach can be used to quickly evaluate and optimize new recommender systems on historical user-generated data. Yet these offline measurements need not translate directly onto the sought-after online results, such as increases in user engagement. This talk will describe the blend of offline and online experimentation we use at Netflix to improve upon our recommendation systems, and will discuss some key challenges and limitations of these approaches that are broadly relevant to the recommender systems field.