Online Evaluation and User Studies

Learning Objectives

- To understand the goals behind user-centered evaluation, and where it complements or replaces offline evaluations
- To become familiar with a variety of mechanisms for user-centered evaluation of recommender systems

Why Evaluate with Users?

- Metrics such as MAE, top-N Precision, diversity, and the like only tell part of the story
  - The real question involves user preference and behavior
  - Don't know how users balance different attributes, different objectives, contexts
  - Need to understand difference between retrieval and recommendation
  - Most of all – user behavior is complex

A Spectrum of Techniques

- Usage Logs
- Polls, Surveys, Focus Groups
- Lab (and online Lab) Experiments
- Field Experiments and Trials

- Other techniques as well … often an area to consult an expert in user studies, HCI, etc.

Usage Logs

- Means for evaluating use of features
  - Success of particular recommenders
    - Very useful with mixed hybrids – track separate success rates
  - Interface issues
- Potential for measuring retrospective accuracy, etc.

Polls, Surveys, Focus Groups

- Can be useful for assessing overall desires, context, usage patterns
  - Be careful: hard to create good surveys
  - Techniques for combining responses, identifying underlying factors
  - Surveys often useful in conjunction with other techniques
- Don't confuse information gathering with selling!

Lab and Online Lab Experiments

- Careful controls, but decontextualized
- Which recommendation list do you prefer?
- Rate these recommendations/lists on the following attributes (familiarity, accuracy, believability, interest, …)
- Likert-scale questions common
- Be careful about question ordering (e.g., general to specific, if goal is to get both)

Field Experiments and A/B Tests

- Beauty of this domain is ability to try variants and measure results …
  - What do you want to measure:
    - Immediate Behavior (recommendations taken, options changed, etc.)
    - Longer-term Behavior (return rate, change in purchases, referrals, …)
    - Subjective Responses (prompts/surveys at strategic times)
- A bit about the culture of massive A/B testing

Take-Away

- No substitute for real user-centered testing
- Need to design tests around goals
  - Different methods can achieve different results
- Need an implementation and users for field experiments and usage logs, but all others can be done with simulated systems

Online Evaluation and User Studies