Conducting RecSys User Studies with POPROX

Robin Burke

Michael Ekstrand

Joseph A. Konstan

POPROX NEWS

What POPROX is for …

I have a great new algorithm that I believe is much better …

I have a new recommendation layout or interface that should make all the difference …

I want to study it on real users, but that's so much work …

- build a system; get human subjects approval; recruit users; keep them long enough to get meaningful results; collect, analyze, release data; …

HELP!

to the rescue!

Who we are …

Here today …

Robin Burke, University of Colorado Boulder
Michael Ekstrand, Drexel University
Joe Konstan, University of Minnesota
Elizabeth McKinnie, University of Colorado Boulder
Sophia Sun, University of Minnesota

Others …

Bart Knijnenburg, Clemson University; Ed Malthouse, Northwestern U.
Daniel Kluver, Karl Higley, Mahamudul Hasan, Zannatun Sristy (Minn.)
Shahan Iqbal (Clemson); Aishwarya Satwani (Colorado),
Sushoban Parajuli (Drexel), Xinyi Li (Northwestern)

AIMS of POPROX

Help researchers …

- Carry out recommender systems experiments with
  - Real, long-term users
  - A comprehensive collection of metrics
  - Best practices for reporting, sharing results

- Change the field
  - Shift away from solely offline evaluation of new algorithms
  - Shift away from one-shot out-of-context user studies
  - Move towards rigorous online experimentation
    - Consistent metrics, open and replicable studies

POPROX Timeline

Idea (v0. approx 2005; v1. 2019)

Planning grant and community consultation (2020-2021)

Project funded (thanks to NSF!) (2023)

Initial development (2023-2024)

First running version (April 2024)

Recruiting participants (Fall 2024)

Recruiting researchers (Fall 2024)

Expected initial external experiments (early 2025)

The Newsletter

POPROX NEWS

Atlantic City mayor and his wife plead not guilty to beating their daughter

The first couple of New Jersey's Atlantic City have pleaded not guilty to beatings and other abuse of their teenage daughter

Read more

Officials work to protect IV supplies in Florida after disruptions at North Carolina plant

Federal officials are working to move IV bags out of the path of Hurricane Milton, which is threatening another manufacturer of IV fluids even as hospitals nationwide are still reeling from disruptions caused by flooding at a large factory in North Carolina

Read more

Ex-NFL Media journalist says he's reached a settlement with the league on discrimination lawsuit

What Does POPROX Offer?

- Agile, Iterative, and Responsive Platform
  - Many capabilities will be developed in response to the first need
- Recommender algorithm and interface studies
  - Selection and order of news articles
  - Changes to text headers and slugs (short summaries)
  - Layout, structure
- User surveys and behavioral data
  - Open, click through, opinion, self-reported behavior
- Why run a study with POPROX?
  - Real users,
- Ability to use POPROX data prior to / without running an experiment
  - Note: for user data protection and content copyright protection, all data access requires a free license

Restrictions and Limitations

Technical Limitations

- Limited by email newsletter interaction – little opportunity for interaction or detailed feedback (planning to seek funding for a v2 app/web interfaces)
- Currently content is all fairly neutral (Associated Press); open to incorporating other content where providers will agree to its use

Researcher Terms and Conditions

- No direct access to participants or personally-identifiable information, no open text responses, no harmful deception, no pornography, no advertising
- News content may not be used to train models exported from POPROX experiments (can validate the model concept, but trained model is for POPROX experiment only)
- News content may not be re-served by others, attributes must be preserved, some other limitations on use of content (e.g., no taking out of context for disparagement)

POPROX Open Research Model

Our Commitment to Open Research

- Open-Source Code
  - We welcome contributions of enhancements back to the platform
- Wide Access to Data (through free license to protect participants/IP)
- Platform for Disseminating Results

Our Expectations for Researchers

- Publish Experiment Manifest on POPROX Site for Others
- Publish Experiment Dataset (after research publication) on POPROX Site

POPROX NEWS

Example Studies

Three example studies …

Recommender Only

- New personalized recommender based on participant history, items, produces a recommendation list (top-k) and uses rest of infrastructure.

Interface Only

- Variation of content presentation, e.g., adding explanatory tags to help participant understand why they would want to read particular content.

Combined Interface and Recommender

- Restructure newsletter to provide more “important” content that does not match user preferences – updates recommender and layout/interface to explain/promote the content.

POPROX NEWS

Architecture

POPROX Architecture Highlights

- Cloud-based (AWS) platform
  - Both our platform and your experiment code live in AWS
  - Bring-your-own-compute (in our availability zone)
- Figure
  - Us
  - Recommender
  - Data (training + article text)

POPROX Platform <> Recommender Interface

POPROX requests recommendations via **HTTP API**

Sends (as JSON):

- User profile data (history + topic preferences)
- Candidate articles (previous day's news)

Expects (as JSON)

- List of articles with display information (headline, blurb, image)

Recommender controls item display

Our code open — start by forking the repository

POPROX Measurements

Behavior measurements

- Article clicks + derived metrics
  - Article CTR
  - Newsletter CTR / hit rate
- Full impression log for click context
  - MRR of first clicked articles
  - NDCG/RBP

Perception measurements

- Regular user surveys with validated instruments to measure satisfaction, novelty, diversity, etc.

Offline measurements

- Standard code includes offline metric support with MIND (and eventually POPROX training data)
- Use full suite of LensKit metrics
- Additional content metrics (RBO, etc.)