Click-Based Recommender Evaluation

Katja Hofmann
Microsoft Research

Anne Schuth
ISLA, University of Amsterdam

Alejandro Bellogin
Universidad Autónoma de Madrid
(work completed while at CWI)

Maarten de Rijke
ISLA, University of Amsterdam

Overview

Recommender System (RS) evaluation has long focused on *explicit* feedback, where users rate specific items. However, much more data can be collected when considering *implicit* feedback, such as clicks.

The question we focus on is: **what is the relationship between RS evaluation using explicit and implicit feedback?**

Experiments

- Given evaluation with explicit feedback, can we predict online performance (measured on implicit feedback)?
- We compare representative RSs: random baseline (RND), popularity-based (ItemPop), item-based (IB), matrix factorization (MF), user-based (UB).
- Compare evaluation with explicit and implicit feedback:
  - **explicit feedback** – nDCG and Precision on explicit labels
  - **implicit feedback** – clickthrough rate (CTR) with the user models: no-bias, examination and browsing models with logarithmic and quadratic discount (exam-{log,quad}, browse-{log,quad}), and cascade model with low and high stop probability (cascade-{low,high}).
- Tasks: (1) rank RS and check agreement, and (2) predict "online performance" (based on implicit feedback) from evaluation with explicit ratings on MovieLens 1M data (with 80-20 test / train split).

Results

- As expected, implicit and explicit evaluation agree well when assumptions agree well (e.g., precision@10 and CTR with no-bias).
- The match between assumption on user behaviour and explicit evaluation really matters – if assumptions are violated, the wrong RS can be preferred / deployed.

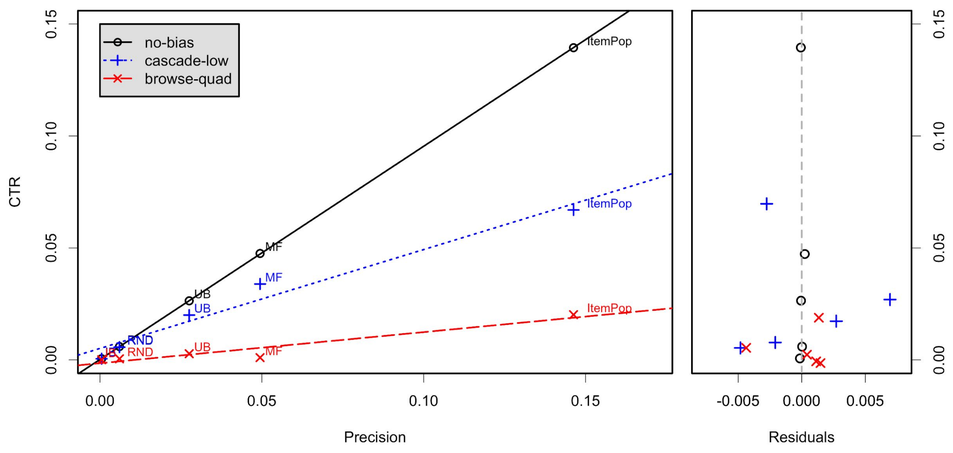
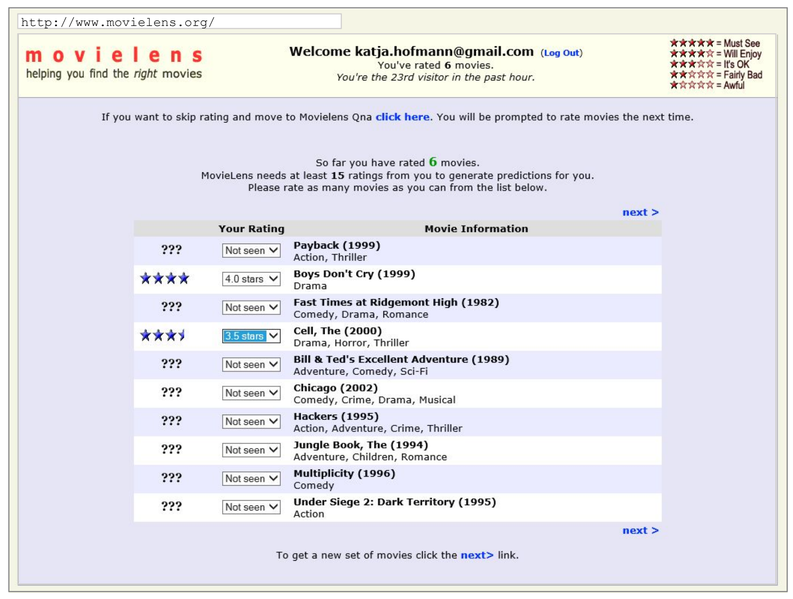


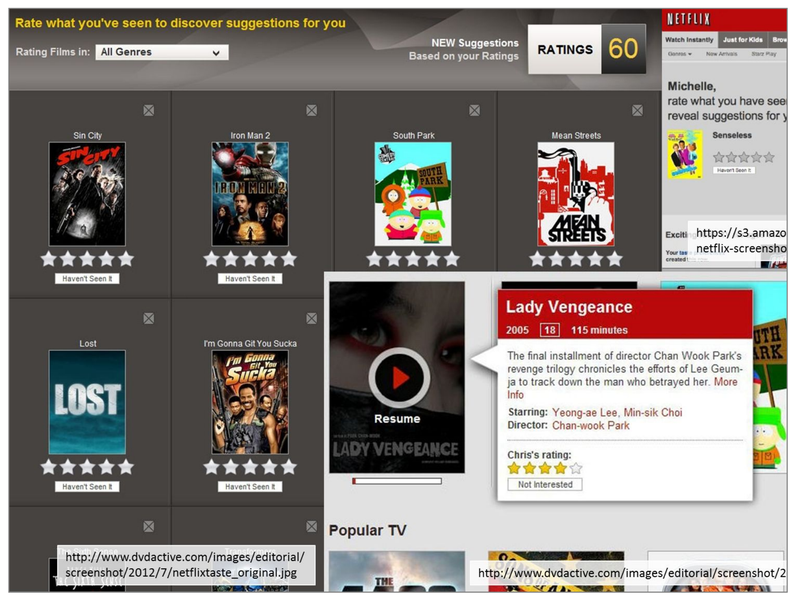
Figure 2: Overview of results. Precision and CTR scores for the tested RS are shown for selected user models. Predicting online performance from explicit feedback works well when assumptions about user behavior are met, but can prefer the wrong RS other-wise.

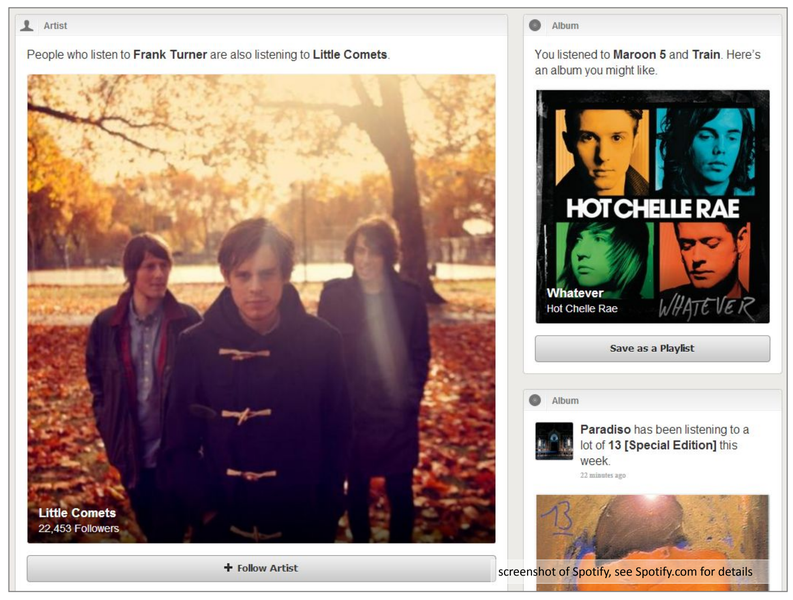
Open questions

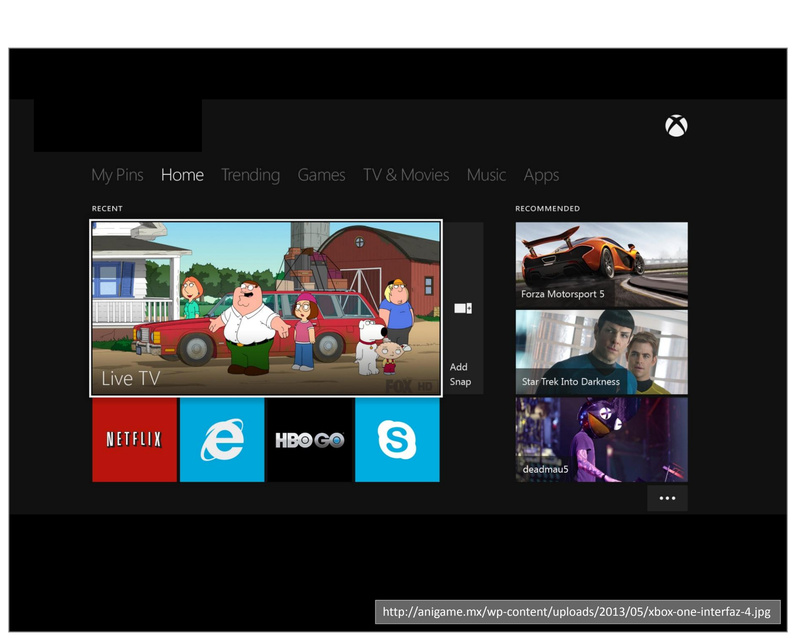
- How should explicit labels be used for RS development before implicit feedback can be collected?
- Can we obtain any guarantees on performance with implicit labels, given offline evaluation with explicit labels?
- How can we deal with missing explicit labels?

Figure 1: Interfaces for collecting explicit and implicit feedback for different RS.









Microsoft Research

ISLA

CWI Centrum Wiskunde & Informatica

LiMoSINe

Katja Hofmann, Anne Schuth, Alejandro Bellogin, & Maarten de Rijke: Effects of Position Bias on Click-Based Recommender Evaluation. ECIR 2014, pp. 624-630.