Offline vs. Online Evaluation in Voice Product Search

Amir Ingber, Liane Lewin-Eytan, Alexander Libov, Yoelle Maarek and Eliyahu Osherovich

Amazon Research - Alexa Shopping

CIKM GLARE workshop, Oct. 2018
Intelligent Personal Assistants [TheWebConf 2018 keynote]

• Meta layer of intelligence
  • Sits on top of other services and applications
  • Performs actions or serves content using services/apps to fulfill the user’s intent
  • Natural language interface
  • Relies on
    • machine learning, AI, speech recognition, natural language understanding, dialog management, ranking, inference, personalization, etc..

• Major IPAs in the Market

  - Siri (2011)
  - Google Assistant (2012)
  - Cortana (2014)
  - Alexa/Echo (2014)

- Limited information flow into smartphones/devices with typing/touch
- People can speak up to 4 times faster than they can type
- Speech is expected to replace touch/typing as the primary input form
  - By 2020 50% of all searches will be voice searches (comScore)
  - By 2020 about 30% of searches will be done without a screen (Mediapos)

- Music
- Information
- Smart home
- Games
- Shopping
- …
“Alexa, buy…”
“Alexa, reorder paper towels”
Voice Product Search

“Alexa, buy dog food for puppies”

“Ok. The top search result is Purina ONE SmartBlend Healthy Puppy Formula Dry Dog Food"
Voice Product Search & Offline eval.

• New paradigm, drastically differs from Web domains
• Results are spoken: fewer results, less information
• Positive shopping actions - mainly performed on the first result

Problem: Offline evaluation
• Widely used for training/evaluation
• Essential in voice shopping, as being a new habit
• Relies on a dataset with relevance judgements, typically derived from history
• Voice shopping data cannot be used “as is” for training
Offline Vs. Online Evaluation

- Lack of correlation between offline and online demonstrated by two experiments.
- Run over one week of voice shopping traffic

1. New ranking model
   - Online A/B testing
   - Ranking Offline results

2. Random ranking of first 5 results
   - Online A/B testing
   - Ranking Offline results
Gap Between Offline and Online

Metric: Mean Reciprocal Rank of converted product \( MRR = \frac{1}{|Q|} \sum_{i=1}^{\lfloor |Q| \rfloor} \frac{1}{rank_i} \)

* MRR Received some criticism lately, N. Fuhr, “Some Common Mistakes In IR Evaluation, And How They Can Be Avoided”, SIGIR Forum 2018

• New Model (MRR)
  • Significant decrease in offline (-15%)
  • Slight increase in online (+1%)

• Random
  • Huge decrease in offline (-43.5%)
  • Some decrease in online (-8.2%)
Observations

- New Model: negative in offline setting, but positive in true online setting
- Random experiment: extreme bias for the first position

Traditional log-based offline evaluation cannot be used directly as proxy for online experiments in voice product search.
Research Questions

- Can we still use logs?
- Manual golden sets?
- Can log-based data be de-biased?
- Use data from random experiments?
We Are Hiring!

It's Day One.
We Are Hiring!