Offline vs. Online Evaluation in Voice Product Search

Amir Ingber, Liane Lewin-Eytan, Alexander Libov, Yoelle Maarek and Eliyahu Osherovich
Alexa Shopping, Amazon Research
{ingber, lliane, alibov, yoelle, oeli}@amazon.com

1 BACKGROUND

Intelligent voice assistants such as Amazon Alexa, Google Assistant, and Apple Siri have been recently gaining popularity. One emerging usage of such assistants is shopping. Being able to restock toilet paper after using the last roll, or ordering garlic, while adding the last clove in your pasta sauce, is a new habit that people are acquiring. In the traditional usage scenario, customers issue a product search query by voice, and get back a list of candidate products on which they can take some actions such as add-to-cart or buy.

Voice product search introduces a new paradigm and drives users' behavior to drastically differ from other domains with closed collections such as Web Mail or Web Product search. As the output is spoken, customers are exposed to fewer results, with much less information. Positive shopping actions, such as buy and add-to-cart, are much more frequent on the first result than in other domains.

Before a ranking model is pushed into production, a common practice is first to evaluate it offline. Offline experiments are much easier to conduct than online experiments and therefore allow for faster iteration in algorithmic improvements. They also act as a safeguard in the hope that they catch defective models before they are being tested on real users, even if in limited numbers. Offline evaluation relies on a dataset with relevance judgments, however, in our context, most judgments are associated with the first result presented. Such datasets are typically derived from historical search logs, which includes search queries, associated results, and actions taken by real users on these results.

In this workshop, we would like to argue that traditional search evaluation methods cannot be used "as is" in voice product search. We will show that log-based offline experiments do not sufficiently correlate with online results to be valuable. This has also been demonstrated recently by Carterette et al. [1] in other domains. Besides, voice shopping still being a new habit, online experiments might be riskier than in other environments such as Web search, as negatively affected users might not try the experience again. We hope to discuss with other attendees the need to invent new types of offline experiments that would be less sensitive to the display order, and of online experiments when data is relatively scarce.

2 OFFLINE VS. ONLINE EVALUATION

We verified the lack of correlation between offline and online evaluation through two experiments that were run over one week of voice shopping traffic. In the first experiment, we evaluated the performance of a new ranking model using an offline and online validation process. In both processes, a new model was compared to the model used in production (referred to as *control* model). In the offline process, historical data was re-ranked using the new model. In the online process, the new model and control model were run in parallel in an online environment, over different parts of the traffic.

In the second experiment, we evaluated the performance of a random algorithm, which randomizes the order of the top five results ranked by the control model. Again, the random algorithm was evaluated in an offline and online process as described before.

As an evaluation metric, we used Mean Reciprocal Rank (MRR)[1], defined as $\frac{1}{k}$, where k, in our context, is the rank of the purchased/added-to-cart item. When evaluating a new algorithm on past data, improvement can only come from ranking the purchased items higher than their original place.

The results of the first experiment are presented in the first row (New model) in Table 1, showing that in the offline evaluation, the new model led to a 15% decrease in MRR, while in the online evaluation, the new model achieved a performance slightly better than control model (+1% MRR).

The results of the second experiment are presented in the second row (Random) of Table 1. As can be seen, the offline evaluation led to a 43.5% decrease in MRR, while in the online evaluation, the decrease in MRR was only 8.2%. This shows a clear bias towards first position results, which should be taken into account in the learning and evaluation processes [2].

The second experiment highlights the extreme bias for the first position result in voice product search, which directly impacts the validity of using historical data in offline experiments. A new ranking algorithm might seem negative in an offline evaluation, but positive in a true online setting as shown in the first experiment. More generally, based on these experiments, we make the case that traditional log-based offline evaluation methods cannot be directly used as a proxy for online experiments in voice product search.

Experiment	offline	online
New model	-15%	+1.0%
Random	-43.5%	-8.2%

Table 1: Relative MRR difference vs. the control model.

3 RESEARCH QUESTIONS

We would like to encourage the community to explore new research directions for adequate training and evaluation of voice product search, which we see as a growing area of investigation. Examples of hard research questions include:

- Can log-based data be leveraged when users are still learning a new medium and their behavior changes fast?
- Can manual golden sets enrich log-based data sets?
- Can log-based data be de-biased?
- Can we leverage data from random experiments?

These questions are just a few in an emerging domain where both customers and researchers need to adapt to a new paradigm and a new user experience.

REFERENCES

[1] Ben Carterette and Praveen Chandar. 2018. Offline Comparative Evaluation with Incremental, Minimally-Invasive Online Feedback. In *SIGIR'18*. 705–714.

[2] X. Wang et al. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *WSDM'18*. 610–618.

[3] Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (2018), 32–41.

[1] MRR is widely used as a common evaluation metric in IR, however, has lately been subject to criticism [3]. We complete our evaluation by also considering the inverse of Mean First Relevant metric, as specified in [3], leading to a similar trend as in Table 1.