Learning to Rank with Selection Bias in Personal Search

Xuanhui Wang [1] Michael Bendersky [1]

[1]Google Inc.

Research Gap

The paper addresses the research gap in effectively leveraging sparse click-through data for learning-to-rank in personal search. Existing click models typically require large quantities of clicks per query-document pair, which is **not feasible in personal search** due to personalized corpora and information needs. Furthermore, obtaining explicit relevance judgments in personal search is difficult due to privacy and the dynamic nature of personal data. The paper also identifies a novel selection bias problem in this context, where queries with clicks are under-sampled in a biased manner, affecting the learning process.

Main Contributions

- Introduces the **selection bias problem** in learning-to-rank for personal search.
- Proposes several **bias estimation models**, including a **query-dependent (generalized) model** that does not require large amounts of click data.
- Presents a novel **unbiased offline evaluation methodology** based on randomized result sets.
- Validates the models through large-scale **offline and online experiments** on a real personal search engine.

Method Overview

The key method used is **inverse propensity weighting** to address the selection bias. The paper estimates the propensity score (the probability of a query appearing in the clicked data) and uses its inverse to weight the loss function during learning-to-rank. Different models for estimating these inverse propensity weights are proposed: a **global bias model** based on position, a **segmented bias model** based on query segments (email labels), and a **generalized bias model** using **multi-class logistic regression with query-dependent features** to predict position bias. **Result randomization** is employed to collect unbiased data for bias estimation.

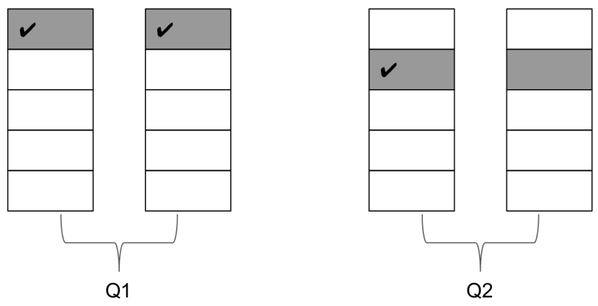

Figure 1: An illustration of selection bias in click data. The shaded documents are the relevant ones. A check mark means the document is clicked.

Figure 1. Selection bias in click data.

Key Findings

- Accounting for **query-dependent selection bias** in learning-to-rank significantly improves search effectiveness (MRR and CTR) in online experiments with a large-scale personal search engine compared to a baseline without bias correction.
- **More fine-grained bias models** (segmented and generalized) lead to further improvements over a global bias model, with the segmented model showing significant improvement over the global model in online A/B testing.
- The proposed **unbiased offline evaluator provides a sound methodology** for evaluating ranking functions on randomized data, addressing the limitations of standard offline metrics in the presence of selection bias.

Strengths and Weaknesses

Strengths:

- **First** to comprehensively study selection bias in personal search ranking.
- Uses **real-world large-scale deployment** with randomized experiments.
- Develops a novel offline evaluation strategy that **overcomes bias** from using regular click data.

Weaknesses:

- The offline evaluation showed **no statistically significant difference** between the segmented and generalized models, potentially due to the limited size of the evaluation data.
- Evaluations limited to **single-click queries** and **personal search scenarios**—not generalized to multi-click or broader search settings.

Future Directions

- Evaluating the applicability of these methods to **web search**.
- Extending the framework to handle **multiple clicks per query**.
- Investigating effective features for **bias estimation in different application domains**.
- Exploring **cheaper and less intrusive** methods for collecting randomized data.
- Improving the **data utilization** of the unbiased offline evaluator.

Core Equation

Inverse Propensity Weighting for Empirical Loss:

$$L_S(f) = \frac{1}{|S|} \sum_{Q \in S} \frac{P(Q)}{\hat{P}(Q)} \cdot l(Q, f) = \frac{1}{|S|} \sum_{Q \in S} w_Q \cdot l(Q, f)$$

Where $\hat{P}(Q)$ is known as the propensity score of Q.

List of position bias prediction methods

Name	Method Description
NoCorrection	No bias correction is applied. This serves as our baseline.
Global	The bias is estimated for each position globally.
Segmented	The bias is estimated for each position per segment.
Generalized	The bias is estimated for each position per query using logistic regression.

Example

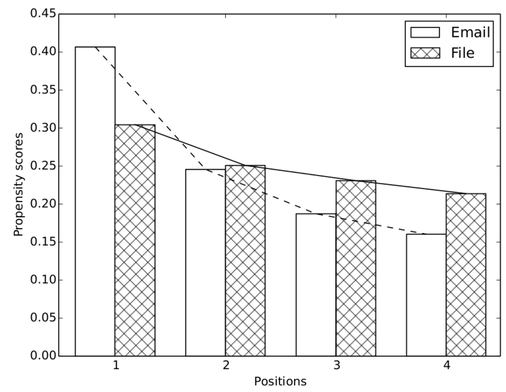

Figure 2: The position bias propensity scores for user emails and cloud storage files.