Training On-Device Ranking Models from Cross-User Interactions in a Privacy-Preserving Fashion

Marc Najork
Google LLC, 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA
najork@google.com

ABSTRACT

Personal search is concerned with surfacing content relevant to an information need (as expressed by a query) from a user's personal information repository. Since personal corpora are typically much smaller than public ones (particularly the web), recall is more of an issue. Moreover, since documents are not shared among users, cross-user interaction signals (such as co-clicked results for identical or similar queries) cannot be leveraged in a straightforward manner. When limited to a single user, interaction signals are typically too sparse to be useful as labels or as features in learned ranking functions.

Bendersky et al. [3] recently described a methodology for leveraging user interactions in the form of clicked search results in a way that allowed them to aggregate interactions across the entire user base of a personal search service, by projecting both queries and documents into a shared, dense feature space, and training a ranking function on these features using result clicks as relevance judgments. Using clicks as relevance labels requires accounting for the inherent selection bias in click logs, which can be measured through short-lived result randomization experiments on a portion of users [7, 12] or learned jointly with the ranking function [2, 13].

In the past several years there has been a lot of interest in training machine-learned models in a federated fashion, suitable for on-device training and inference [8]. To prevent leakage of personal information, one can leverage ideas from differential privacy, where noise is added to any training record proportional to the sensitivity of that record [5]. Several recent works have studied the topic of learning with differential privacy in a federated setting [1, 4, 6]. In the same time period there has been tremendous interest in the IR community on privacy-preserving IR, manifested by three workshops and two tutorials; see https://privacypreservingir.org for a good overview.

Can we adapt the ideas from on-device learning using privacy-preserving federated shared models to personal information retrieval? Fundamentally, ranked retrieval from personal corpora involves three types of data, all of which are privacy sensitive: documents (e.g. files, photos, messages, music, videos etc); queries (including query reformulations and refinements over the course of a search session), and implicit feedback such as click and attention signals [9]. Much of the existing work in privacy-safe federated learning has focused on marrying stochastic gradient descent-style optimizations with differential privacy (see e.g. [11]). Some portions of the framework for jointly estimating position bias and training a ranking function [13] (e.g. using gradient boosted decision trees as a ranker) fit nicely into such a framework; other aspects (e.g. enforcing k-anonymity thresholds on query and document n-grams) will require new research. The same holds true for other search improvements that involve learning, such as improving recall through synonym expansions trained from query reformulations or result co-clicks [10].

We hope that this abstract will inspire researcher in Information Retrieval to explore this exciting new frontier of privacy-safe on-device personal search.

DESIRES 2018, August 2018, Bertinoro, Italy

REFERENCES

[1] Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, and Li Zhang. 2017. On the Protection of Private Information in Machine Learning Systems: Two Recent Approches. In *30th IEEE Computer Security Foundations Symposium (CSF).* 1–6.
[2] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W. Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR).* 385–394.
[3] Michael Bendersky, Xuanhui Wang, Donald Metzler, and Marc Najork. 2017. Learning from User Interactions in Personal Search via Attribute Parameterization. In *10th ACM International Conference on Web Search and Data Mining (WSDM).* 791–799.
[4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2016. Practical Secure Aggregation for Federated Learning on User-Held Data. *CoRR* abs/1611.04482 (2016). arXiv:1611.04482
[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *3rd Theory of Cryptography Conference (TCC).* 265–284.
[6] Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially Private Federated Learning: A Client Level Perspective. *CoRR* abs/1712.07557 (2017). arXiv:1712.07557
[7] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *10th ACM International Conference on Web Search and Data Mining (WSDM).* 781–789.
[8] Jakub Konecný, Brendan McMahan, and Daniel Ramage. 2015. Federated Optimization: Distributed Optimization Beyond the Datacenter. *CoRR* abs/1511.03575 (2015). arXiv:1511.03575
[9] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).* 113–122.
[10] Cheng Li, Mingyang Zhang, Michael Bendersky, Hongbo Deng, Donald Metzler, and Marc Najork. 2018. Embedding-based Synonyms for Personal Search. (2018). Under submission.
[11] Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP).* 245–248.
[12] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).* 115–124.
[13] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *11th ACM International Conference on Web Search and Data Mining (WSDM).* 610–618.