VIRLab: A Platform for Privacy-Preserving Evaluation for Information Retrieval Models

Hui Fang  
Department of Electrical and Computer Engineering  
University of Delaware  
Newark, DE USA  
hfang@udel.edu

ChengXiang Zhai  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL USA  
czhai@illinois.edu

ABSTRACT

Information retrieval (IR) has been a highly empirical discipline since the very beginning of the field. The development and study of any novel techniques such as retrieval models always require extensive experiments over multiple representative data collections. Traditionally, IR evaluation relies on the use of publicly available data, so researchers often download the collections and conduct the evaluation on their servers. However, this would not be a favorable (or even possible) solution to evaluation over the proprietary data due to various privacy concerns. In this paper, we discuss one potential solution to the privacy-preserving evaluation (PPE) for IR models. We first briefly introduce the VIRLab system, and then discuss how to extend the system to enable a controlled data-centric experimental environment for evaluation over proprietary data.

Categories and Subject Descriptors: H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness)

General Terms: Experimentation

Keywords: virtual IR lab; privacy-preserving evaluation; PPE; data-centric evaluation

1. INTRODUCTION

Information retrieval (IR) is an empirical discipline. The research progress achieved in this field is closely related to careful and thorough evaluation over representative data collections. For example, when developing a new algorithm such as a retrieval function, it would be necessary to compare its performance with those of the state of the art retrieval functions on multiple representative data collections. Since traditional data collections are often publicly available, researchers could download the collections and conduct the evaluation on their own servers, as shown in the left plot of Figure 1. We refer such an evaluation paradigm as algorithm-centric evaluation since the evaluation happens at the site of algorithms and data are moved there.

Although the current evaluation practice works well with the publicly available data collections, it would not be able to support the evaluation over proprietary data, which can not be easily shared due to various privacy concerns. As a result, only a very small number of researchers who have the access to these proprietary data collections are able to conduct experiments, which makes it impossible for other researchers to reproduce the results. Clearly, the current practice has the serious problem of not being able to reproduce experimental results over private data collections.

Due to the empirical nature of the discipline, it is always essential to evaluate IR algorithms with real applications involving real users, and thus almost always raise the issue of privacy protection. Clearly, it is important to study how to improve the reproducibility of IR research and enable controlled experiments on proprietary data while preserving the privacy of the collections.

In this paper, we propose a novel privacy-preserving evaluation (PPE) paradigm, which is data-centric. Instead of conducting the evaluation at the sites of algorithms, the new evaluation paradigm moves the algorithms to the data and conducts evaluation at the sites of the data, as illustrated in the right plot of Figure 1.

To support the proposed PPE paradigm, it would be necessary to develop an infrastructure that enables users to upload the code of their methods, evaluates the uploaded codes and returns the evaluation results to the users. We first discuss the challenges of building such an infrastructure, and then explain how to leverage the recently developed Virtual IR Lab (VIRLab) system to overcome the challenges.
users to implement retrieval functions, evaluate the functions over the provided data collections and then analyze the evaluation results when necessary. We now describe how to leverage the VIRLab system to solve the three challenges discussed in the previous section.

**Dynamic code generation for algorithm uploading:** The VIRLab system currently allows users to implement a retrieval function through a Web form by combining statistics provided through a list. After that, the implementations are converted and embedded to C/C++ codes, and the codes are then compiled and executed. The process of code conversion is achieved by a customized dynamic code generator [1]. We propose to adapt the similar strategies to more general scenarios such as allowing users to upload their own code that following the conversions required by the dynamic code generators.

**Modularized evaluation infrastructure:** To enable the evaluation of individual IR system components, we propose to modularize the evaluation infrastructure. Such a design could allow users to evaluate the effectiveness of each component. Although the current VIRLab system allows only the customized implementation of retrieval functions, we plan to open up other components and allow users to upload or implement their own methods. In fact, such a design could enable a more controlled experiment set up for privacy-preserving evaluation.

**Multi-level privacy-preserving result delivery:** The VIRLab system provides a leaderboard for each data collection, which displays the evaluation results for well-performed retrieval functions. Moreover, it also allows users to see the performance for each query and compare the performance of their methods with a baseline method. All the above information would not contain much sensitive information since only the evaluation results are reported and no information about the private data has been revealed. It is clear that such a strategy can protect privacy well, but might not provide lots of useful information for the users to analyze and figure out how to revise the retrieval functions to improve the performance. Since not every private collection has the same level of privacy concerns, it would be necessary to identify multiple privacy-preserving levels and decide how to return results accordingly. For example, we could anonymize query terms with their IDs and display the statistics for each term. If the data collection has less restriction, we could consider to display the actual terms or phrases without revealing the actual content in the data. Finally, constructing diagnostic evaluation collections [2] would be another way of diagnosing the problems of a retrieval function without giving out private information.

4. CONCLUSIONS

In this paper, we propose a novel *data-centric PPE* evaluation infrastructure. The basic idea is to move the evaluation process from “algorithm sites” to “data sites”. Its unique advantage is to enable the evaluation over proprietary data while preserving the privacy. We identify three challenges and propose to leverage the VIRLab system to solve them.

5. REFERENCES

