Privacy-Preserving IR 2016: Differential Privacy, Search, and Social Media

Hui Yang
Georgetown University, USA
huiyang@cs.georgetown.edu

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Li Xiong
Emory University, USA
lxiong@emory.edu

Charles L. A. Clarke
University of Waterloo, Canada
claclark@plg.uwaterloo.ca

Simson L. Garfinkel
NIST, USA
simson.garfinkel@nist.gov

ABSTRACT

Due to lack of mature techniques in privacy-preserving information retrieval (IR), concerns about information privacy and security have become serious obstacles that prevent valuable user data to be used in IR research such as studies on query logs, social media, and medical record retrieval. In SIGIR 2014 and SIGIR 2015, we have run the privacy-preserving IR workshops exploring and understanding the privacy and security risks in information retrieval. This year, we continue the efforts of connecting the two disciplines of IR and privacy/ security by organizing this workshop. We target on three themes, differential privacy and IR dataset release, privacy in search and browsing, and privacy in social media. The workshop include panels with researchers from both fields on these three themes, as well as invite industry speakers for real-world challenges. The goals of this workshop include (1) bringing together the two research fields, and (2) yielding fruitful collaborations.

Keywords

Privacy-Preserving Information Retrieval

1. MOTIVATION

Information retrieval and information privacy/security are two fast-growing computer science disciplines. Information retrieval provides a set of information seeking, organization, analysis, and decision-making techniques. Information privacy/security defends information from unauthorized or malicious use, disclosure, modification, attack, and destruction. The two disciplines often appear as two areas with opposite goals: one is to seek information from large amounts of materials, the other is to protect (sensitive) information from being found out. On the other hand, there are many synergies and connections between these two disciplines. For example, information retrieval researchers or practitioners often need to consider privacy or security issues in designing solutions for information processing and management, while researchers in information privacy and security often utilize information retrieval techniques when they build the adversary models to simulate how the adversary can actively seek sensitive information. However, there have been only limited efforts to connect the two important disciplines.

Due to lack of mature techniques in privacy-preserving information retrieval, concerns about information privacy and security have become serious obstacles that prevent valuable user data to be used in IR research such as studies about query logs, social media, tweets, session analysis, and medical record retrieval. For instance, the recent TREC Medical Record Retrieval Tracks are halted because of the privacy issue and the TREC Microblog Tracks could not provide participants with a standard testbed of tweets for system development. The situation needs to be improved in a timely manner.

In addition, with the emergence of online social networks and the growing popularity of digital communication, more and more information about individuals is becoming available on the Internet. While much of this information is not sensitive, it is not uncommon for users to publish sensitive information online, especially on social networking sites. The availability of this publicly accessible and potentially sensitive data can lead to abuse and expose users to stalking and identity theft. User privacy is at risk whenever owners of publicly available network resources can be traced and recorded, and their user activities can be mined and inferred. The digital ecosystem involves not only such threats to personal privacy but also promising cures. For example, information-theoretic protocols may decompose each user's query into several sub-queries, ensuring they include no information about the user's intent.

All these motivate us to propose this "privacy-preserving IR" workshop in SIGIR.

2. PAST WORKSHOPS

Since 2014, the privacy workshop has been running in SIGIR. The first privacy-preserving IR workshop (PIR 2014) [5] received active participation. There were three keynotes given in the workshop. Most workshop attendees agreed that

SIGIR '16 July 17-21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2917763

we need to continue to collaborate on this topic, and to seek both general and specific solutions to the privacy and security issues in various IR tasks. The second workshop (PIR 2015) [7] started to receive solutions to the problems that we identified in the previous year. It featured a keynote on differential privacy and a heated debate on "privacy would be resolved if the search is not personalized". Both years' workshop reports can be found in SIGIR forum too.

3. WHAT'S NEW

This year, the workshop organizers formed a team spanning across IR, Privacy, and the government sector. We focus on forming IR and privacy research bridges and partnerships and thus organizing panels by themes. For each theme, we invite researchers from IR and researchers from Privacy/Security to form pairs. The panel discussion not only focus on presenting problems but more on seeking solutions and achieving a mutual understanding of the research problem. Among many challenges in privacy-preserving IR, this year we focus on three major themes: *differential privacy for IR data release*, *privacy-preserving search/browsing* and *privacy and social media*.

4. THEMES

The three main themes of this year including the following.

- Differential privacy for IR data release [1, 2]: How to preserve privacy when releasing internal data. For instance, web query logs and medical records could not be shared without privacy concerns. Another challenge is how to anonymize the datasets and make sure that they can be shared with a certain degree of privacy guaranteed while preserving the utility of the data. We focus on the latest technology of differential privacy in particular.

- Privacy preserving search/browsing [3]: How to recognize and communicate privacy concerns to users of search engines. It is important to develop effective and efficient solutions to protect users' privacy in information retrieval applications while they are searching or surfing the net. We also plan to discuss softwares such as TunnelBear[1].

- Privacy and social media [4, 6]: New information retrieval and natural language processing technologies are needed to quickly identify components and/or attributes of a user's online activities on social media that may reduce the user's privacy, and warn users about their vulnerability on the Web. For instance, authorship attribution.

5. ACTIVITIES AND SCHEDULE

This full day workshop includes invited speakers, panels, short contributed talks and a poster session. It ends with a breakout session. The activities are scheduled as the following:

09:00-09:30 Welcome and Introduction.
09:30-10:30 Invited speaker #1 (Differential privacy)
10:30-11:00 Coffee
11:00-11:45 Contributed talks (session #1)
11:45-12:30 Posters flowing into Lunch
12:30-13:30 Lunch
13:30-14:30 Invited speaker #2 (Industry)
14:30-15:15 Contributed talks (session #2)
15:15-16:30 Panel and breakout planning
16:30-17:30 Breakouts (research challenges)
17:30-18:00 Summary of breakouts

[1]https://www.tunnelbear.com/

6. WEBSITE

The website of the workshop can be found at privacypreservingir.org.

7. ACKNOWLEDGMENTS

This research was supported by NSF grant CNS-1223825, DARPA grant FA8750-14-2-0226, and NSF student travel grant. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor. We also express our special thanks to our student organizer, Sicong Zhang.

References

[1] L. Fan, L. Bonomi, L. Xiong, and V. Sunderam. Monitoring web browsing behavior with differential privacy. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, 2014.

[2] M. Gaboardi, E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu. Dual query: Practical private query release for high dimensional data. In *ICML 2014, Beijing, China, 21-26 June 2014*.

[3] W. Jiang, L. Si, and J. Li. Protecting source privacy in federated search. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 761–762. ACM, 2007.

[4] W. Luo, Q. Xie, and U. Hengartner. Facecloak: An architecture for user privacy on social networking sites. In *CSE '09 - Volume 03*, pages 26–33, Washington, DC, USA, 2009. IEEE Computer Society.

[5] L. Si and H. Yang. Pir 2014 the first international workshop on privacy-preserving ir: When information retrieval meets privacy and security. *SIGIR Forum*, 48(2):83–88, Dec. 2014.

[6] L. Singh, G. H. Yang, M. Sherr, A. Hian-Cheong, K. Tian, J. Zhu, and S. Zhang. Public information exposure detection: Helping users understand their web footprints. In *ASONAM 2015, Paris, France*.

[7] H. Yang and I. Soboroff. Privacy-preserving IR 2015: When information retrieval meets privacy and security. In *SIGIR'15 , Santiago, Chile, August 9-13, 2015*.