Privacy-Preserving IR 2015: When Information Retrieval Meets Privacy and Security

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

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

ABSTRACT

Information retrieval (IR) and information privacy/security are two fast-growing computer science disciplines. There are many synergies and connections between these two disciplines. However, there have been very limited efforts to connect the two important disciplines. On the other hand, due to lack of mature techniques in privacy-preserving 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, tweets, and medical record retrieval. We propose this privacy-preserving IR workshop to connect the two disciplines of information retrieval and information privacy and security. We look forward to spurring research that aims to bring together the research fields of IR and privacy/security. Last year, the first privacy-preserving IR workshop focused on mitigating privacy threats in information retrieval by novel algorithms and tools that enable web users to better understand associated privacy risks.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

Keywords

Privacy-Preserving Information Retrieval

1. MOTIVATION

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. An adversary can digitally "stalk" a victim (a Web user) and discover as much information as possible about the victim, either through direct observation of posted information or by inferring knowledge using simple inference logic.

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 of 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 very limited efforts to connect the two important disciplines.

In addition, 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 example, the recent TREC Medical Record Track halted over concerns about privacy in nominally de-identified data, and the TREC Microblog Track does not directly distribute tweets in order to support the right of Twitter users to make private or delete their tweets and accounts. The situation needs to be improved in a timely manner. All these motive us to propose this "privacy-preserving IR" workshop in SIGIR.

Co-located with SIGIR 2014, the first privacy-preserving IR workshop (PIR 2014) [3] received active participation. Most workshop attendees agreed that 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. This year, we continue the efforts with a set of new focuses.

2. THEME AND PURPOSE

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

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767857.

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.

Privacy and security has been discussed in a few recent IR conferences and workshops. The SIGIR 2012 "Community IR Evaluation using Private Data Collections" workshop discussed issues how private data such as medical records can be shared for open IR evaluations. The CIKM 2013 "Channeling the Deluge: Research Challenges for Big Data and Information Systems" Panel Session also briefly addresses the privacy issue in information systems. There are also privacy- and security-related workshops that less related to IR but more related to data mining, artificial intelligent, and machine learning. For instance, Computer and Communication Security (CCS) 2013 workshop on "Artificial Intelligence and Security". However, their focus is not mainly for IR and is very different from the proposed workshop. As far as we know, there have been very limited efforts to connect the two important disciplines of information retrieval and information privacy and security.

This workshop aims to spurring research that aims to bring together the research fields of IR and privacy/security, and mitigate privacy threats in information retrieval by constructing novel algorithms and tools that enable web users to better understand associated privacy risks. We believe this workshop in SIGIR can provide valuable opportunities to explore the connection between information retrieval and information privacy and security.

2.1 List of Questions

The workshop addresses several important topics that connect information retrieval, information privacy and security:

- Protecting User Privacy in Search, Recommendation and Beyond [1]: much damage can be caused as users can be identified in AOL query log data and Neflix log data, it is important to develop effective and efficient solutions to protect users' privacy in information retrieval applications.
- Dataset Distribution and Evaluation: How does privacy affect IR test dataset distribution and evaluation? For instance, web query logs and medical records could not be shared without privacy concerns to the public or the researchers. How to anonymize the datasets and make sure that they can be shared with a certain degree of privacy guarantee while at the same time preserves the utility of the data?
- Information Exposure Detection : new information retrieval and natural language processing technologies are needed to quickly identify components and/or attributes of a user's online public profile that may reduce the user's privacy, and warn one's vulnerability on the Web.
- Novel Information Retrieval Techniques for Information Privacy/Security Application: new information retrieval, evaluation, or machine learning techniques need to be designed that fit the practice of applications in information privacy and security.
- Private Information Retrieval Techniques for Enabling Location Privacy in Location-Based Services [2]: data about a user's location and historical movements can potentially be gathered by a third party who takes away the information without the awareness of the service providers and the users, how location-based services and recommender systems interact with Location Obfuscation techniques and other Privacy-Enhancing Technologies.

3. KEYNOTES

Speaker: Dr. Li Xiong, Emory University.

Title: Making Private User Data Accessible for Information Retrieval Research: Data Sharing with Differential Privacy

Abstract: Almost a decade has passed since the infamous AOL data leak. Up till today, data privacy concerns continue to prohibit valuable user data such as query logs and web browsing sessions to be shared and used for Information Retrieval (IR) research. As a result, lack of large scale datasets is still one of the major barriers facing academic IR researchers. This talk will give an overview of the recent developments in building and sharing statistical and synthetic datasets based on private data with the rigorous differential privacy guarantee. Then we will focus on techniques we have developed for sharing sequential data under differential privacy, addressing challenges such as high dimensionality and high correlations, which are also common characteristics in user behavior data. Empirical studies using real-world web browsing data will demonstrate the feasibility as well as challenges of applying differential privacy on user behavior data for IR research.

4. ACTIVITIES AND SCHEDULE

9:00–9:15am Welcome A brief welcome which illustrates the goal of this workshop and the schedule for the day. All participants introduce themselves, their research focus, and the specific interest in the topic of the workshop.

9:15–10:30am Keynote Talks

10:30–12:30pm Morning Presentations We have a morning presentation session for high quality papers that are accepted by the workshop.

12:30–1:30pm Lunch Break Participates have their lunch, can discuss about the morning talks on their own.

1:30–4:00pm Group Discussion There will be breakout groups on different themes to have further discussion in this section.

4:00–5:00pm Bring it All Together At the end of the day, we will get everyone together to discuss about the invited talks and posters.

5:00–5:15pm Conclusion The organizers will summarize the discussions and conclude the day.

5. WEBSITE

Our website is located at `privacypreservingir.org`.

References

[1] 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.

[2] A. Khoshgozaran and C. Shahabi. Privacy in location-based applications. chapter Private Information Retrieval Techniques for Enabling Location Privacy in Location-Based Services, pages 59–83. Springer-Verlag, Berlin, Heidelberg, 2009.

[3] 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.