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

Luo Si
Purdue University, USA
lsi@purdue.edu

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

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, sessions, and medical record retrieval. This privacy-preserving IR workshop aims to spur research that brings together the research fields of IR and privacy/security, and research that mitigates privacy threats in information retrieval by constructing 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 & THEMES

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. 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. However, there have been limited efforts to connect the two disciplines.

A few instances of IR practices have raised concerns about privacy and security. One famous example is the AOL query log release that raised a great deal of controversial discussions about how IR researchers can use available online materials to advance their research while preserving users' privacy and data security. In fact, 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. 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.

Some research themes of the workshop include:

- Protecting User Privacy in Search, Recommendation and Beyond [2]: 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;
- Information Exposure Detection [4]: new information retrieval and natural language processing technologies are needed to quickly identify components 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 or machine learning techniques need to be designed to fit the practice of applications in information privacy and security;
- IR Techniques for Enabling Location Privacy in Location-Based Services [3]: 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 recommender systems can interact privacy-enhancing technologies such as Location Obfuscation;
- Private Information Retrieval (PIR) and Data Sharing [1]: new cryptographic protocols are needed to safeguard the privacy of IR users; they allow clients to retrieve information from shared datasets while completely hiding the identity of the retrieved data.

2. REFERENCES

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *Journal of ACM*, 45(6):965–981, Nov. 1998.

[2] W. Jiang, L. Si, and J. Li. Protecting source privacy in federated search. In *SIGIR*, pages 761–762, 2007.

[3] A. Khoshgozaran and C. Shahabi. Privacy in location-based applications. Springer-Verlag, 2009.

[4] W. B. Moore, Y. Wei, A. Orshefsky, M. Sherr, L. Singh, and H. Yang. Understanding site-based inference potential for identifying hidden attributes. *PASSAT*, pages 570–577, 2013.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2600737.