PIR 2014
The First International Workshop on 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

On July 11th, 2014 the First International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (PIR 2013) [1, 2] was held as part of the SIGIR 2014 conference in Gold Coast, Australia. Three invited keynotes were given by Prof. Douglas W. Oard, Prof. Christopher W. Clifton, and Prof. David D. Lewis. There were 6 full papers and 3 short papers presented in the presentation sessions. Finally, there were breakout discussions on determining the future directions of these interdisciplinary fields of IR and privacy/security for our community.

1 Introduction

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

2 Invited Keynotes

2.1 Search Among Secrets: Separating the wheat from the buzzsaw

The opening keynote was given by Douglas W. Oard (University of Maryland, College Park, USA) on "Search Among Secrets: Separating the wheat from the buzzsaw".

A fundamental assumption of nearly all information retrieval research is that content that should not be shown to the user should simply not be included in the collection that is indexed. This assumption breaks down, however, when it is not practical to separate what we want (the "wheat") from what needs to be protected from disclosure (the "buzzsaw" that can ruin your day). In this talk, Douglas motivated the problem from four perspectives: multi-level security in enterprise search, scholarly access to personal papers in archival institutions, requests by citizens for access to government records, and withholding of privileged content in the "discovery" process that arises from civil litigation in some jurisdictions. For each of the last three cases, Douglas described a research project that sheds some light on requirements, challenges, and capabilities. Douglas then concluded the talk by offering thoughts on task design, evaluation design, and system design that may help to move us closer to being able to address this increasingly important grand challenge of search among secrets.

2.2 (Semi)Private Information Retrieval: A Discussion of Privacy Risks and Relaxation

The second keynote in the morning was given by Christopher W. Clifton (Purdue University, currently on leave at the National Science Foundation, USA) on "(Semi)Private Information Retrieval: A Discussion of Privacy Risks and Relaxation".

Private Information Retrieval started with a very strict premise: Nobody should learn anything about the query or what is retrieved. The results are not encouraging: Even simple Document ID based retrieval becomes impractical. But we have also seen that openly disclosing queries and results can lead to clear privacy violations.The solution must lie in between these extremes. This talk looked at some past work that takes very different approaches to protecting privacy, and recent debates of privacy risks and harm. Christopher laid out foundations for discussing privacy requirements in information retrieval. The goal was to lead to a discussion among the workshop participants that will identify opportunities for future research in Privacy-Preserving Information Retrieval.

2.3 Privacy in Creating Text Classifiers for Electronic Discovery

After lunch, David D. Lewis (David D. Lewis Consulting, USA) gave a keynote lecture on "Privacy in Creating Text Classifiers for Electronic Discovery".

A large civil lawsuit, particularly in the United States, can require the categorization of billions of enterprise documents to determine which need to produced to opposing parties. Text classifiers, whether created manually or trained by supervised learning, have become the only practical approach to this task, with positive predictions usually reviewed by attorneys prior to production. The adversarial setting poses challenges, however, since producing parties view both interactive development of search queries and labeling of training data as risking the revelation of sensitive information. David discussed the approaches that have developed for meeting these challenges in e-discovery, make connections with the literature on privacy-preserving IR and data mining, and suggested some directions for research that would be of great benefit in e-discovery.

3 Papers Presented at the Workshop

We requested the submission of both long papers (6 pages) and short papers (2 pages) to be presented orally in the workshop. We accepted a total of 6 long papers and 3 short papers. During the workshop, each long paper was given a slot of 15 minutes, while each short paper was given a slot of 10 minutes, for oral presentation.

3.1 Long Papers

The first long paper was *HEALTH+Z: Confidential Provider Selection in Collaborative Healthcare P2P Networks* by Sergej Zerr, Odyseas Papapetrou, Elena Demidova. They provided a privacy preserving IR work for the healthcare domain. Many real world applications in the healthcare domain would gain a substantial advantage from sharing and search technologies available for P2P infrastructures if these technologies could provide required confidentiality guarantees. Currently, DHT-based indexes which are typically applied for effective and efficient information sharing and retrieval in P2P networks do not offer sufficient confidentiality for the patient data in a healthcare network and medical document archives. In this paper they discussed the challenges involved in securing patient data stored in a DHT-based index and discuss initial solutions to address these challenges.

The second long paper was *Privacy-Preserving Important Passage Retrieval* by Luís Marujo, José Portêlo, David Martins de Matos, João P. Neto, Anatole Gershman, Jaime Carbonell, Isabel Trancoso and Bhiksha Raj. They presented a privacy preserving method for important passage retrieval that relies on creating secure representations of documents. Their approach allows for third parties to retrieve important passages from documents without learning anything regarding their content. They use a hashing scheme known as Secure Binary Embeddings to convert a key phrase and bag-of-words representation to bit strings in a way that allows the computation of approximate distances, instead of exact ones. Experiments show that their secure system yield similar results to its non-private counterpart on both clean text and noisy speech recognized text.

The third long paper was *Benchmarking the Privacy-Preserving People Search* by Shuguang Han, Daqing He and Zhen Yue. This paper focus on the privacy issues in people search. They proposed simulating different privacy settings with a public social network due to the unavailability of privacy-concerned networks. Their study examines the influences of privacy concerns on the local and global network features, and their impacts on the performance of people search. Their results show that: 1) the privacy concerns of different people in the networks have different influences. People with higher association (i.e. higher degree in a network) have much greater impacts on the performance of people search; 2) local network features are more sensitive to the privacy concerns, especially when such concerns come from high association peoples in the network who are also related to the querying user.

The fourth long paper was *Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data* by Yi Song, Daniel Dahlmeier, Stephane Bressan. This paper worked on anonymizing location data. They proposed a modification-based anonymization approach that is based on shorting the trajectories to reduce the risk of re-identification and information disclosure. Empirical, experimental results on the anonymized dataset show the decrease of uniqueness and suggest that anonymization techniques can help to improve the privacy protection and reduce privacy risks, although the anonymized data can't provide full anonymity so far.

The fifth long paper was *User Comment Analysis for Android apps and CSPI Detection with Comment Expansion* by Lei Cen, Luo Si, Ninghui Li, and Hongxia Jin. This paper investigated the security and privacy issues during analysis of mobile app comments. Specifically, they collected a dataset of comments from Google Play, and proposed a two dimensional label system to describe those Security/Privacy Issues (CSPI) within it. A supervised multi-label learning method utilizing comment expansion was adopted to detect different types of CSPI described by this label system. Experiments on the collected dataset show that the proposed method outperforms the method without the comment expansion.

The last long paper was *Personalisation of Web Search: Exploring Search Query Parameters and User Information Privacy Implications - The Case of Google* by Anisha T. J. Fernando, Jia Tina Du, and Helen Ashman. This paper worked on the personalization of web search. It highlights the need to explore search query parameters and determine their impact on personalisation. This is a first step in exploring the mechanisms of personal data collection and how personalised search uses personal data, which subsequently impacts the information privacy of users. It was found that location parameters have more impact on personalisation than the parameter 'pws' that switches personalisation on or off. Hence, it is important to undertake further research that investigates the impact of other types of search query parameters, their contribution towards search personalisation and their impact on user information privacy.

3.2 Short Papers

The first short paper was *VIRLab: A Platform for Privacy-Preserving Evaluation for Information Retrieval Models* by Hui Fang and ChengXiang Zhai. They discussed one potential solution to the privacy-preserving evaluation (PPE) for IR models. They first briefly introduced the VIRLab system, and then discuss how to extend the system to enable a controlled data-centric experimental environment for evaluation over proprietary data.

The second short paper was *On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records* by Timothy Gollins, Graham McDonald, Craig Macdonald, and Iadh Ounis. They discussed challenges that arise from two stages in the archiving digital government records, which information retrieval research can address: the selection / appraisal of appropriate records to archive, and the review of those records to endure that no sensitive information is released. They also suggest tentative solutions for sensitivity review.

The last short paper presented was *Increased Information Leakage from Text* by Sicong Zhang, Hui Yang, Lisa Singh. They offered insight into the textual information leakage problem, focusing on current relevant research and potential areas of synergy between the information retrieval community and the privacy community. They analyze example publicly shared data and discuss the types of data that can be extracted, the methods used for extracting them, and the implications for individuals who share personal information.

4 Breakout Discussions

After the paper presentations, we had one hour breakout discussions. Here are some topics discussed during the breakout.

1. What are the most important issues for further research in privacy-preserving information retrieval? New search algorithms?

   Dr. Douglas W. Oard raised the question about the definition of privacy itself. It seems that it is quite a relative concept and is used from diffrent places with different meanings in reseach areas. And different people may have differenct sense of what privacy is. It could means confidentiality, anomynity or something else. People agreed that this is quite a problem if the definition of privacy itself is not clear.

2. How does privacy affect IR test dataset distribution and evaluation? For instance, TREC could not share datasets in the tasks of medical record retrieval, micro blog retrieval?

   The medical record retrieval in TREC has been closed due to privacy issues. For a traditional IR platform, privacy issues exists in many phases. First, the privacy of user query; second, the privacy of the search log kept by the engine; last but not least, the privacy issue from the release of these data for research dataset, as in TREC.

3. How to connect the research with government regulations like EU's "right to be forgotten" and President Obama's office's "Big Data and Privacy" white paper?

   Dr David D. Lewis refered the "right to be forgotten" to be just the "right to get a entry in database deleted" and what was on internet may not be really "forgotten", and the age of the ability to erase things has gone. This point receiced a lot of feedback from different point of view. For the users, they do need to right to demand to withdraw privacy information from internet, even it is only to control the damage. Whether the private information is kept by some company or publicly available is still different.

4. How is sensitivity different from privacy?

   A point is presented as that private information is about you, but involves you, while sensitivity is about you. But its more contextual. This piont wasn't agreed by all.

5. Other topic and questions discussed.

   People also discussed about the following points:

   - whether IR techniques can help to solve or detect sensitive data?
   - Is it a good idea to move algorithms to the data?
   - Tasks related in retrieval: how can we hide the real users? How can we identify/protect the sensitive information of the documents. How can we obsfucate the queries and documents representation?

5 Conclusion

Overall, the workshop was an productive and successful event with active participation from the community. In the workshop, we confirmed that the security and privacy challenges are widely existing in various IR tasks. From an overall standpoint, three inspiring keynotes brought different perspectives on the privacy preserving IR field. Specifically, this workshop also brought together a diversified set of privacy-preserving IR works on healthcare data, passage retrieval, people search, location data anonymization, mobile app comment analysis, personalization of web search, IR model evaluation, sensitivity review, and information leakage from text. Finally, participants discussed on some major concerns of the field. Most of our 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. We expect to see more relevant works to be produced by our participants and together to create the future of the privacy preserving IR field.

References

[1] L. Si and H. Yang. Privacy-preserving ir: When information retrieval meets privacy and security. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 1295–1295, New York, NY, USA, 2014. ACM.

[2] L. Si and H. Yang, editors. *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37th Annual International ACM SIGIR conference, PIR@SIGIR 2014, Gold Coast, Australia, July 11, 2014*, volume 1225 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.