Search Among Sensitive Content

Graham McDonald[1(✉)] and Douglas W. Oard[2]

[1] University of Glasgow, Glasgow, UK
graham.mcdonald@glasgow.ac.uk
[2] University of Maryland, College Park, MD, USA
oard@umd.edu

Keywords: Sensitive information · Sensitivity-aware IR · Information leakage

Information retrieval (IR) systems provide access to large amounts of content, some of which may be personal, confidential, or otherwise sensitive. While some information protection is legislated, e.g., through the European Union's General Data Protection Regulation (GDPR) or disclosure exemptions in Freedom of Information Acts; other cases are regulated only by social expectations. Information retrieval research has traditionally focused on finding indexed content. However, the increased intermixing of sensitive content with content that can properly be disclosed now motivates research on systems that can balance multiple interests: serving the searcher's interest in finding content while serving other stakeholders' interests in appropriately protecting their sensitive information.

If the content requiring protection were marked, protecting it would be straightforward. There are, however, many cases in which sensitive content must be discovered before it can be protected. Discovering such sensitivities ranges in complexity from detection of personally identifiable information (PII), to automated text classification for sensitive content, to human-in-the-loop techniques for identifying sensitivities that result from the context in which the information was produced.

Once discovered, IR systems can use the results of sensitivity decisions to filter search results, or such systems can be designed to balance the risks of missing relevant content with the risks of disclosing sensitive content. Optimising sensitivity-aware IR systems' performance depends on how well the sensitivity classification works, and requires development of new evaluation measures. Moreover, the evaluation of such IR systems requires new test collections that contain (actual or simulated) sensitive content, and when actual sensitive content is used secure ways of evaluating retrieval algorithms (e.g., algorithm deposit or trusted online evaluation) are needed. Where untrusted data centres provide services, encrypted search may also be needed. Many of these components rely on algorithmic privacy guarantees, such as those provided by k-anonymity, L-diversity, t-closeness or differential privacy.

This tutorial will introduce these challenges, review work to date on each aspect of the problem, describe current best practices, and identify open research questions. Resources and outputs from the tutorial can be found on the tutorial website: https://search-among-sensitive-content.github.io.

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, p. 730, 2021.
https://doi.org/10.1007/978-3-030-72240-1