Searched Subject’s Privacy in Search Engines: Concerns and Mitigations

Technology Review – CS 410 Text Information Systems 2021 - kkhare2@illinois.edu

Introduction

Debates and discussions on privacy in Search Engines context primarily focus on the privacy of end user using the Search Engine, and how their details along with multiple other meta fields (like time of search, location of search, pre and post searches, IP address, browser details etc.) are routinely collected when these users do searches looking for information on the internet [Zimmer 2008], and how this collected information can be mined, profiled, and traded for commercial purposes [Pariser 2011], and can compromise the privacy of these end users via identification, monitoring and surveillance.

The lesser discussed aspect of privacy in Search Engines pertains to the privacy of the subject (individual or entity) who is searched, and whose information is getting indexed, stored, and surfaced to the end users for purposes known or unknown to the subject, and with or without the consent of the subject. This technology review paper explores this “searched subject” aspect of privacy in search engines, the key concerns around it, and suggests two naïve approaches on how the retrieval systems underlying search engines can take into account subject’s private content in the data and surface the results appropriately or not surface them at all to end users.

Privacy for Searched Subjects

Search Engine users at times use the search technology to look for information on individuals (or entities like an organizations). These individuals are the subject of the search and considering they may not have given explicit consent for storing, indexing, and surfacing personal data on them, ethically it may be unfair to them if this personal information about them on internet gets used to violate their privacy and is detrimental to their well-being. Also, as the information about individuals on internet is increasing at an ever-expanding rate not just via online forums and social networking sites public pages, but also as more and more public and private organizations digitize their record-keeping and use internet as the service delivery and communication mechanism, the chances of individual’s personal information becoming public (inadvertently or otherwise) and accessible via Search Engines is increasing significantly. The individuals in these cases have little control on how information about them can be acquired by Internet users, which in turn has implications for personal privacy [Tavani 2005].

Two hypothetical examples to illustrate these implications are mentioned below:

1. A candidate getting rejected by a university admission committee after one of the admission committee’s conservative members searched for the candidate online and noticed photographs of the candidate drinking in a casino party. Even though there is no way to establish causality that the photographs surfaced by the Search Engine led to

candidates' rejection, it is hard to ignore that the committee member may have got influenced by the photographs, even though they have no bearing on candidate's application.

2. A person contributing to LGBT rights support organization via donating money, with this organization publishing the list of all donors on their website (known or unknown to the person). Person's colleagues in a new organization he/she joined searching person's name in an online search engine, noticing that the person contributed to the LGBT rights support organization, and carrying the impression that person is gay/lesbian. How this impression will impact the person or whether the person is gay/lesbian is immaterial – the key here is person's colleague were able to find information about the person via a Search Engine which the person may or may not have been comfortable in sharing with them.

In both the above examples, the subjects' information on internet was surfaced by Search Engines to people for whom that information was irrelevant, and the subject may not even be aware that this information is on the internet or that the people around the subject may have easy access to this this information, thus impacting their interactions and transactions with the subject i.e., Subject was treating that information as private in this situation, and subject's privacy was violated to some degree.

Search Engines accountability for violation of Searched Subject privacy

The key arguments which a Search Engine company can make to emphasize that they have minimal role in violation of searched subject's privacy are:

1. They are not responsible for the content on the internet and are just neutrally indexing and surfacing the content already created

2. They are only indexing the publicly available data on the internet, and if a user's private data is on internet, then it is user's liability and not of the Search Engine company

The first argument does not hold much ground considering Search Engine companies do treat different sets of data on internet differently for indexing and surfacing purposes. A very stark example of this is pornography content or copyrighted content, which may be lying on the internet and accessible by Search Engine crawler bots, but still the Search Engine either don't surface them to end users or surface them with appropriate cautions, adhering to the law of the land and ethical considerations. In addition, most Search Engines nowadays are highly user centric and produce customized results e.g., a search on phrase like "Climate Change" may yield different results list depending on where the user has done the query from, and what are users political and professional affiliations (e.g., a company executive or an environmental activist may get different results), thus neutrality of search results is not guaranteed by design [Pariser 2011].

The second argument even though seems legitimate, has two counterarguments to take in consideration:

1. Is it only the subjects who puts their own data on the internet?

   The answer to this question is a clear No. Adding sensitive information (sensitive in specific context) about a person can be a totally benign activity e.g., two friends going out for drinks, and one taking their selfie and uploading it on internet (assume a public page), or an organization publishing the list of all its donors; or it can be done on purpose to malign the subject e.g. a disgruntled employee releasing the private emails from his manager in public domain to shame the manager or publishing false/fabricated stories about the manager showing him in bad light. As soon as the data is available on public pages, Search Engines will index it and start surfacing it to other users.

2. How do we define "public" data for a person? What if the data about the person is public but sensitive in specific contexts? Should Search Engines treat all public data for a person equally?

   Based on the research on this topic [Tavani 2005], "Some forms of personal information enjoy normative protection via policies and laws because they involve data about persons that is either sensitive or intimate, or both. This kind of personal information can be referred to as Non-Public Personal Information (or NPI). This could include information about a person's finances and medical history. In contrast Public Personal Information (or PPI), is personal in nature, but considered to be neither intimate nor confidential (e.g. which school a person goes to or what car does a person own." So far most of the concerns regarding individual's privacy have been limited to NPI data sets, but as the volume and variety of PPI data generated has increased exponentially, and collection & analysis technologies have evolved significantly, PPI data itself can result (intentionally or unintentionally) in impacting the well-being of the subject, and hence how PPI data sets get used by Search Engines becomes more critical and highly relevant to the "problem of protecting privacy in public [Nissenbaum 1998]".

Considering the above, there is a clear case that Search Engines companies should take pro-active steps to limit the violations of privacy of searched subjects, both due to Non-Public Personal Information and due to Public Personal Information.

Mitigations Search Engines can implement to protect privacy of Searched Subjects

This section suggests two naïve technical approaches which Search Engines can use to limit the violations of privacy of search subjects.

1. Non-Public Personal Information (or NPI) in web page content can be largely identified based on sensitive identifiers (e.g., SSN, National ID, Passport Number, Driver's License Number etc.) and NPI key words/phrases (e.g., "Medical Records", "Laboratory Reports"). Search engine indexers can use Regex expression to confirm and count the presence of sensitive identifiers and can do the phrase matching to confirm and count the presence

of NPI key words. These counts and presence flags can either be a signal for indexer to not index the document, or else they can be stored in Inverted Indexes, and can be used to not surface NPI documents to public at large. In exigency scenarios (e.g., related with Law Enforcement), these documents can be surfaced to appropriate stakeholders.

2. Public Personal Information (or PPI) is contextually sensitive in nature any may not be amenable to Rule-Based identification as can be the case with NPI. With the advancement in NLP and Text Analysis techniques, machine learning can be used to categorize the documents which are getting indexed in two categories: PPI Positive and PPI Negative, which can be stored in Inverted Indexes. During the search process, PPI Positive documents can be given lower scores (e.g., by adding a constant in the denominator of BM25 scoring function), thus decreasing its rank, and ensuring it doesn't surface in the top results for users doing casual searches on the subject, thus limiting its impact.

Conclusion

Privacy of searched individuals (subjects) is equally important as the privacy of the users doing the search via internet Search Engines. In this technology review paper, we highlight these privacy concerns using hypothetical but realistic examples, and then argue Search Engines have a role to play in ensuring that privacy of searched subjects is not violated thus negatively affecting their well-being. We then differentiate between Non-Public Personal Information (or NPI) and Public Personal Information (or PPI) based privacy concerns and suggest two naïve approaches on how the retrieval systems underlying search engines can consider NPI and PPI data about crawled pages in consideration and limit (or at least reduce) the privacy violations of the searched subjects. During the research done for this technology review, author was unable to verify if any of the mentioned approaches are in use by Search Engines already and if they have been researched in detail, hence the naïve nature of them. Author looks forward to exploring these approaches more and implementing them (if not already done) in future studies on this topic.

References

- Pariser, E., 2011. *The Filter Bubble: What the Internet is Hiding from You*, New York: Penguin.
- Zimmer, M., 2008. "The Gaze of the Perfect Search Engine: Google as an Institution of Dataveillance," in *Web Search: Multidisciplinary Perspectives*, A. Spink and M. Zimmer (eds.), Berlin: Springer-Verlag, pp. 77–99.
- Stanford Encyclopedia of Philosophy (SEP), 2020. Search Engine and Ethics. (https://plato.stanford.edu/entries/ethics-search/)
- Tavani, H.T., 2005. "Search Engines, Personal Information, and the Problem of Protecting Privacy in Public," *International Review of Information Ethics*, 3: 39–45
- Nissenbaum, Helen. (1998). "Protecting Privacy in an Information Age," Law and Philosophy, Vol. 17, pp. 559-596.