Katholieke
Universiteit
Leuven

**Department of
Computer Science**

PRIVACY IMPACT ASSESSMENT REPORT

Personalization Search by Google Search Engine

Sarah Binta Alam Shoilee (r0729856)
Academic year 2018–2019

Contents

1 Introduction

Life of today largely depends on the web search engines; Without an effective one world is quite unimaginable now. Thousands of web documents are available with potential information for any give search query which as a consequence makes it hard to come up with relevant text retrieval. Search algorithms have evolved a lot with time since it's beginning to meet user expectation. Today, Google proves to be the market giant in the particular field which posses over 92.3% of the search engine market share. [1] As a search engine, it successfully manages to fetch useful results based on the search relevancy, ranking the documents to provide most useful information concerning the very topic. Typical search query rate is tens of millions per day and to process all them Google indexes ten to hundreds of millions of web pages each day. [2]
At present, Google not only manages this huge amount of queries everyday, but also process them for future use by introducing many products and services. One of the best features of Google search engine is personalization search. By introducing this feature, Google made the task of searching much more accessible than before by incorporating it with the data from search history and personal records. It was crucial to downsize the amounts of the result generated by a simple query to a reasonable amount that satisfies the user's purpose.
This data integration from different sources surely enhances the search experiences for individual user, but at the same time it raised some privacy concern. In this paper, we are going to asses the privacy issues with Google search engine and possible recommendations to deal with it. At the beginning of the paper, we summarize over the basic mechanism of a search engine to understand our context better.

2 Application Description

2.1 High Level Architecture

To understand the search architecture, we must date back to 1998, when Larry Page and Sergey Brin published their first paper on the anatomy of a large-scale hypertextual Web search engine. The whole proposed architecture stated in Figure-1. The key challenge then was to produce a meaningful result for each query using data-structure efficiently. On that time, to rank a document within first 10 result, which has a high precision Providing a high precision was the ultimate goal. Since then Google search engine has evolved a lot to meet today's need and added many other exciting features apart from the text-based information retrieval. To understand Google's sophisticated search algorithm, it is better to put it in a manner of the first published architecture.To describe further, we are breaking down the entire mechanism into parts:

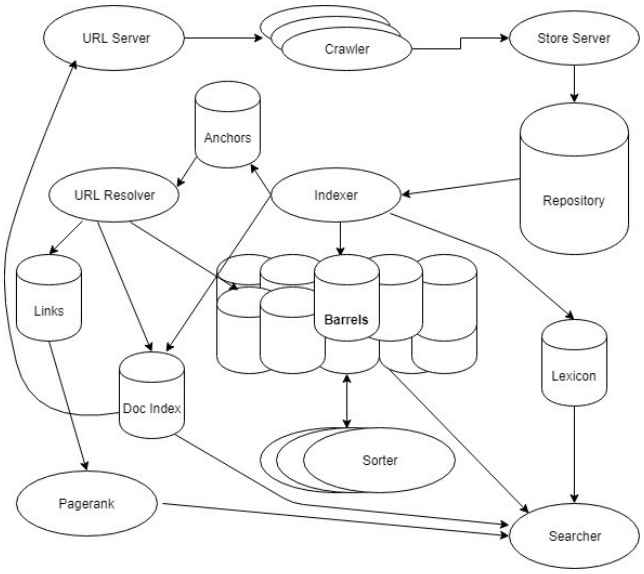

Figure 1: Google High Level Architecture [2]

2.2 Component Details

2.2.1 Web-Crawling

To explain the entire search engine mechanism, URL Server is a good starting point. It sends the list of URLs to the crawler. A web crawler or spider is Google's automated program that is used to organize information from publicly available web pages. Before a search is attempted, it gathers information from hundreds of Billions of web-pages from overall world wide web. Now, Google provides a "WebMaster Tool" to help websites with how their pages will crawl and to design their pages accordingly. [12]
Store-server is responsible for storing the web pages in 'Repository' after compression which has been fetched by the crawler. It is stored using half of its necessary storage by merging all the compressed HTML pages of a single web. [2]
URL Resolver takes the anchor file (a file which contains link of other web pages) and convert the relative URL into an actual URL and put the anchor text to the file it was pointing. It also creates a database of links with its DocID. Every webpage gets its docID when it discovers a new URL. Document Index is fixed length since it contains only fixed length information about each page like status, repository position, checksum etc. For URL and title, it maintains an auxiliary index. [2]

2.2.2 Indexing

Indexer performs the essential task of indexing and parsing. As it gets the web pages from the repository, it uncompressed it. Indexer parses each word of the document and the hypertext link. They record each occurrence of words, word position in the report, font-size, capitalization which is hits. Hits are distributed in a set of 'Barrel' by the indexer. They store the hypertext links to the 'Anchor file' with its relevant information like the source and destination of the link.
Sorter rearranges files within the 'Barrels'. It sorts with WordID, assigned to each word, instead of the DocID like former. By thus, it creates an inverted index.
Lexicon takes the list of wordIDs with its offset in inverted index and produces a new lexicon for searcher replacing the one Index has produced in the earlier stage. Lexicons are nothing but memory-based hash tables. [2] [1]

2.2.3 Page-Rank

PageRank bring a revolutionary change in the queries search result. PageRank decides the rank for each returned website by using the link database for each document. PageRank works by counting the number of links points to this page but not by considering all pointer equally. It also normalizes the amount of the link. In 1998, the PageRank algorithm, used by Google's founder, can be defined by the following equation:

$$PR(A) = (1 - d) + (\frac{PT(T1)}{C(T1)} + ... + \frac{PT(Tn)}{C(Tn)}) \quad (1)$$

where PR(A)= PageRank of page A; T1....Tn= Link to A or refers to & d = damping factor and 0<d<1.[2]
In addition to PageRank, Google, over the years, has added many other secret criteria for determining the ranking of pages on result lists. This is reported to comprise over 250 different indicators, [8][9] the specifics of which are kept a secret to avoid difficulties created by scammers and help Google maintain an edge over its competitors globally. Using PageRank Google is also able to provide personalization in search and makes it impossible for a website to deliberately improved it ranking by misleading. [2] Nowadays Google consider hundreds of factors to a rank website. If any proficient site is referring to one website in a query, chances are the PageRank will be improved.
Many consultancy agencies have been built to help the website to appear more in Google search. To upgrade the page ranking, the basic idea is to use the keyword more often within the page and title. It is presumable that more appearance of the keyword will help website's Google ranking. However, too many occurrences of the keyword cause the page to look suspect to Google's spam checking algorithms. [1]

2.2.4 Searcher

The search algorithm uses the lexicon along with inverted index and PageRank. [2] Over the years, Google has adapted different search algorithm to facilitate users with the result by understanding what they are looking for by interpreting spelling mistake or using the help of natural language programming.
From 2012, Google has started a semantic search. [7] Semantic search not only based on PageRank or indexing but also understand the scientific meaning of the query. Semantic search systems consider various points

including the context of search, location, intent, the variation of words, synonyms, generalized and specialized questions, concept matching and natural language queries to provide relevant search results. [6]
A good number of scientists and engineer are working to improve the user's experience. In 2016 alone, Google made 1600 improvements in their search algorithm. [13] Though it already handled trillions of searches, even now 15% search queries per day appears as they have never seen before. [14] Today Google works like libraries offering millions of books, a travel agent proving transport information or help users by providing anything they need.

3 Stake-holder

In order to collect data, Google states, "We use various technologies to collect and store information, including cookies, pixel tags, local storage, such as browser web storage or application data caches, databases, and server logs." [25] They also mentioned that they share non-identifiable personal data with mass people to show the trend of entire Google services. We can say, the entire population who are benefited by this can be considered to be a stakeholder of Google. To make this data sharing process transparent, Google shares the statistics publicly about the number of requests they receive and their shared amount of information. [26]

3.1 Data Generator

Individual: To facilitate with personalized search result, Google stored the record for previous search queries and browing history. Based on the record, search result varies for every user. They not collect data not only about the users but also any people related to them by tracking down the browsing activity and account information. By inferring E-mail communication, location data and browsing history, it becomes easy to profile a person and even also someone closely related to the user.

3.2 End Users

Individual Google serves users by providing tailored search results. It gathers data from all over the web and benefits the users by proffering the best-suited result. When the system has more and more data about some topic, it becomes easier to pick which document is most related to a given search query. Depending on the number of user visit per document, Google rank it higher and thus serve the next user from previously recorded behaviour.

Government and State Agencies As a significant online enterprise, Government agencies consider Google a massive data source. Google does provide information to the legislative party based on the "good belief" in case of an enforceable government request. They share information for legal reasons upon government request if they have reasonable belief on the parliamentary parties that the information accessed, used, or disclosed only to applicable law, regulation, legal process, or enforceable governmental request. [17] CIA and NSA are two major intelligence agencies have also been reported to be beneficiaries of Google to spy over billions of users. [24]

Corporate Agencies Third parties partners with Google like advertisers, publishers, developers and the right holder may also acquire anonymized data from Google. Google earns its primary revenue from the advertisement over the search engine. So, other organizations apart from Google become beneficiary from the data they collect from users.

4 Privacy Impact Assessment

Given the scope of data sharing via Google search engine, it is important to analyze the privacy risk of it and how Google are dealing with it. It is also crucial to discuss What other things are necessary to cope up with this data emerging world without compromising privacy. The current section describes this in details:

4.1 Privacy Measures by Google

4.1.1 Data Control at Individual Level

According to GDPR, it is imperative that " *the data subject has given consent to the processing of his or her data for one or more specific purposes*" [Article 6 GDPR] [27]. In compliance with the following, Google offers the user control to manage all their data. A user can avail the services of Google in multiple ways. First, the user can log in into their profile and manage privacy according to their manner. Then they can choose the data they want Google to collect and according to that they need to set permission. Google also provides the user option to use their services without prior login or even without creating an account at all. There is also another option for private browsing in Google Chrome's incognito mode. [17]
Google account settings provide good control over data possession too. Users can export their data anytime they want, to create a backup or to see what information Google has in favour of them. Users can also search for data by date or topic. Users are always in power to delete one specific activity or search history or even their entire profile or any previous Google history they wish. Google always provide the user with an option to manage and review their activity. The information that Google collects and use it in user's favor entirely depends on the user and their account management of privacy settings.

4.1.2 Customised Option and Minimize Data Sharing

Along with data control, Google offers its users specific setting of their searching mode. To minimizes the risk of volition of privacy while accessing search in public spaces, user can choose whether he/she wants to stop the personalization mode [18]. Moreover, for advanced manipulation, while accessing Google Search Engine or Chrome Browser, the user will be notified when Google server is setting a cookie in their browser, and the user also can block cookies from other third-party sites; may be from one specific domain or all domains.
Google can also keep track of users' location details by collecting data from GPS, IP Address or any other access point nearby to provide geolocation-based services. However, users can forcefully block location from their device by changing the location share settings. This option is also accessible from the user's Google account location option. [17]
Google shows personalized ads based on the users account activity and their account settings. They apply machine learning to find out on what item the customer will be interested in by monitoring their clicking behaviour on the query result. Users can manipulate the ads she/he will allow Google to show and enable to set their "ad settings" based on their data. However, Google assures that they "don't show personalized ads based on sensitive categories, such as race, religion, sexual orientation, or health." [17]

4.2 Legal Concerns

4.2.1 Search data: classified as personal data

The data Google collects from search engine can also be comprised of sensitive data like financial details, health concern etc. along with other personal keywords. By monitoring search history for quite a long time, it is easy to create a profile of an individual with his or her region, health condition or any other sensitive issue. Some of these data may be classified as private data and can be covered under European General Data Protection Regulation which is in effect from May 25, 2018. To clarify personal data GDPR dictates *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly" or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"* From this definition, we can consider that the data collected from the user's Google account, location data or inferred from previous search which is the primary sources of information for personalization search falls under GDPR. [27]

4.2.2 Data Collection in the Age of Google

The main concern about Google's privacy is the amount of data, they collect. Beginning from the framework like Android, Google has broadened its circle from the web application to various types of gadgets. Moreover, they are likely to approach some third-party application or sites who utilises Google's services in it. Following the trend of Google's data collection, it can be inferred that their plan of action is to gather however as much information about you as could be expected and cross-associate it, so they can endeavor to interface your online persona with your disconnected persona. Some consider this following is only essential to their business,

instead of clients good and termed it as 'Surveillance capitalism'. Under GDPR, it is explicitly mentioned in Purpose limitation principle: Article 5 (1) that "*Personal data shall be adequate, relevant and limited to what is necessary for relation to the purposes for which they are processed* "and "*collected for specified, explicit and legitimate purposes*". [27]

4.2.3 Lack of Transparency

- Google share data with any other organization affiliated to it, law informant agencies or even to the third party advertisement agencies who explicitly accomplices Google's privacy policy to gather data from users' program or gadget for publicizing and estimation purposes utilizing their own treats or comparative advances.
- Google also mentioned that " **We may also collect information about you from trusted partners, including marketing partners who provide us with information about potential customers of our business services, and security partners who provide us with information to protect against abuse.**" [17] There is no further mention about there "trusted partners" and what qualifies as such.
- If someone clicks on an ad resulting from search query, it is possible to share a user's personal information with Google unknowingly. In some event, if you press "tap to call" in any restaurant advertisement provided by Google, they will share your phone number with the restaurant owner without providing further notice.

4.2.4 Unreliable document

To produce a search result, Google crawl over the entire world wide web, mostly from publicly accessible sites. When a search is made using someone's name, Google can present information from some article which may or may not be true according to the data subject. In events, when someone's name shows up in an article, Google's database may record that article if it is openly accessible and show it to other individuals when someone searches for the first person's name. [17] This might sometimes be a case of intrusion of private space. Recently, there was a case raised against Google Spain in European Union's Court of justice by a Spanish man, Mario Costeja González, when his names come up with an auction notice of his house from 1997 which he repossessed later. This case was in demand of the right of delisting this article from Google's Database. [19]

4.3 Social Concern

Anonymous User: In a public server when people are not logged in to his/her respective account, still Google remember the browsing history by assigning a unique identifier on the browser, application or device in use. Google does this to facilitate the user to provide with the best feature of current application depending on the screen setting, operating system or device type. This creates a potential opportunity for the possible breach of privacy on a shared computer. Even if when a user is not logged in, they collect and store data as a unique identifier of the user's device or IP address.

Incognito Browsing: There is plenty of confusion concerning what the "Incognito mode" on Google Chrome does. The majority of people believe that when they go incognito, the user can't be followed. It is a wrong presumption because it is not that Google can't support, Google can follow user even in incognito. The best they do in this circumstance that they chose not to. In any case, this does not imply that some other sites can't follow the user. Whenever a site has visited that tracks and record all the movements of the client even in an ordinary program. It is not quite near to a method of being mysterious on the web. It only forces Google alone to quit following a user.

With immense online popularity and a massive number of users, Google is taking the enormous responsibility of keeping users data private. From search history, it is very much possible to shape actual profile of a person from knowing his language preference to daily activity details with the help of document tracking, location search and previous search records. The whole mechanism is based on user's trust on this online platforms to benefit from digital services. Undoubtedly, it is also a responsibility of Google to act accordingly.

5 Recommendation

In this section, we will sketch recommendations that would help alleviate the privacy concerns that we recognized in the past segment. The given recommendations are divided into three parts which are applicable on both the corporate body and the government in their own respective way.

5.1 Transparency

Data Minimization: Being a tech-giant, Google collects a lot data through all their including search portal. This data collection should be limited to specific to it's need. It is mentioned in the GDPR Article: "Principles Relating to Processing of Personal Data" that "adequate, relevant and limited to what is necessary for relation to the purposes for which they are processed". [27] Google should decide and explicitly mention exactly what amount of data they require to maximize their facility and be clear in words in their privacy policy. The purpose of data collection and data reuse must be clear to the user. In order to make sure that user read and understand their purpose, the corporate should conduct extensive user centric research to protect privacy according to user's need.

Legislative transparency: Government or legislative body also have some responsibility in this regard. They should take control to ensure the transparency and better clarification. Google's Data privacy doesn't address necessary inquiries concerning whether clients can believe that Google's controls or not. There are actually multiple clauses for Data sharing, effective in a different region. Like GDPR, there should be a law for data protection for other regions too.

5.2 Data Protection

Corporate Responsibility: Google possesses a great responsibility because they deal with massive amount of data. Google collects all sort of data from users to make their online experiences better. The data controller should survey data accumulation, stockpiling, preparing works on, including physical safety efforts, to counteract unapproved access to their frameworks. According to GDPR, "The controller shall be responsible for, and be able to demonstrate compliance". [27]

State Responsibility: Like GDPR, there should be a law for data protection for other regions too. Even in the United States, there is no specific national data protection law. There is only "Safe Harbor Principles" which permits US organizations to self-guarantee satisfactory compliance with essential information security standards. In this era of digitization, this is required from National authorities to execute and uphold data protection laws.

5.3 Privacy Protection

Maximum Privacy: Google provides excellent control over the personal data to its users, but it is not entirely helpful. To get into the core of the privacy concern, it is advised to keep the privacy setting at the maximum level by default. It is easy for the users to lose track of control among so many privacy option and leave options open unintentionally. The service provider may ask later the users to enable those settings just by describing the effect of it or to experience more effective performances. For instance, Google account gives you numerous security controls, yet the default for the vast majority of them is to enable Google to gather the best amount of data about you as would be prudent. On such security, control is "Do Not Track" option. By using this, you can explicitly tell one website not to track. If a user wants to enable this, he must dig into his advanced Chrome setting which is not very easy for a non-tech savvy user.

Enhancing User-Interface Design: It is presumable that users are not willing to go through all privacy option to be secured. Considering that setting one single page of privacy options is not all Google can offer to ensure privacy. There should be better user interface design tailored to data subject and to make sure that whatever concern might raise from certain data sharing with the policy against, it does not get ignored. Intensive user-interface design research must be conducted to implement such system.

Digital Awareness: It is critical to offer digital education to understand how much data individual are generating every day. The government should provide direction on the understanding of information security requirements for users who have never really thought about the reality of the data that Google collects; a visual reminder could be a valuable wake-up call.

6 Conclusion

Today, Google is assumed to be one of the leading online services provider. With the introduction of Google Analytics service, it gained further acclaim as well as allow them to access data even if one does not use Google product of services. It is evident that 75% website uses Google analytics to measure their performance. [23] For daily life, it is impossible to avoid such service in fear of privacy. Instead, it should be maintained in some privacy-preserving manner. A big company like Google and their decision towards prioritizing privacy have the potential to shift the culture of online behavior.
Google has done numerous researches to find out the optimal pointer for choosing a search technique. In the end, they found out what matters the most is the search, not the personal information. [20] Jonathan Zittrain, Harvard law professor, expressed doubt to the degree to which personalization search changes Google query result, saying that "the effects of search personalization have been light. [21] A research was conducted by the Nottingham University Researchers between logged in user and a control group to find out the effect of search customization. According to their findings, only 11.7% results make a difference due to personalization. [22] This result disapproves the whole data collecting tendency for Google's personalization search.

References

[1] En.wikipedia.org,
Google. [online] Available at: https://en.wikipedia.org/wiki/Google [Accessed 1 Nov. 2018].

[2] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), pp.107-117.

[3] Ehrlinger, Lisa; Wöß, Wolfram (2016). "Towards a Definition of Knowledge Graphs" (PDF).

[4] En.wikipedia.org. (2018). Knowledge Graph.
[online] Available at: https://en.wikipedia.org/wiki/Knowledge_Graph [Accessed 11 Nov. 2018].

[5] Dewey, Caitlin (May 11, 2016). "You probably haven't even noticed Google's sketchy quest to control the world's knowledge". The Washington Post. Retrieved December 10, 2017.

[6] John, Tony (March 15, 2012). "What is Semantic Search?". Techulator. Retrieved July 13, 2012.

[7] 7. Efrati, Amir (March 15, 2012). "Google Gives Search a Refresh". The Wall Street Journal. Retrieved July 13, 2012.

[8] 8. "Corporate Information: Technology Overview". Google. Retrieved November 15, 2009.

[9] 9. Levy, Steven (February 22, 2010). "Exclusive: How Google's Algorithm Rules the Web". Wired.com. Archived from the original on April 16, 2011.

[10] 10. Albanesius, Chloe (August 10, 2012). "Google to Demote Sites With 'High Number' of Copyright Complaints". PC Magazine. Ziff Davis. Retrieved December 9, 2017.

[11] 11. "Block explicit results on Google using SafeSearch". Google Search Help. Google. Retrieved December 9, 2017.

[12] 12. Google.com. (2018). How Google Search works — Crawling indexing.
[online] Available at: https://www.google.com/search/howsearchworks/crawling-indexing/ [Accessed 4 Nov. 2018].

[13] 13. Google.com. (2018). How Google Search works — Search algorithms.
[online] Available at: https://www.google.com/search/howsearchworks/algorithms/ [Accessed 5 Nov. 2018].

[14] 14. Google.com. (2018). How Google Search works — Useful responses.
[online] Available at: https://www.google.com/search/howsearchworks/responses/

[15] Culliss, G. (March 25, 2003). PERSONALIZED SEARCH METHODS. Overland Park, KS (US).

[16] Google Personalized Search. (2018).
Retrieved from https://en.wikipedia.org/wiki/Google_Personalized_Search

[17] Google Privacy Terms. (2018).
Retrieved from https://policies.google.com/privacy?hl=enfootnote-unique-id

[18] Ludwig, Amber. "Google Personalization on Your Search Results Plus How to Turn it Off". NGNG. Retrieved August 15, 2011. Google customizing search results is an automatic feature, but you can shut this feature off.

[19] Travis, A., Arthur, C. (2018). EU court backs 'right to be forgotten': Google must amend results on request.
Retrieved from https://www.theguardian.com/technology/2014/may/13/right-to-be-forgotten-eu-court-go

[20] Grankvist, Per (2018). How Technology Makes It Harder to Understand the World (1 ed.). United Stories Publishing. pp. 179–180. ISBN 9163959909.

[21] Weisberg, Jacob (June 11, 2011). "Is Web personalization turning us into solipsistic twits?". Slate. Retrieved February 11, 2018.

[22] "A Better Understanding of Personalized Search". Briggs, Justin. Retrieved on Dec 1 2014.

[23] Privacy Issues with The Internet's Most Popular Websites - Search Encrypt Blog. (2019). Retrieved from

[24] https://choosetoencrypt.com/news/privacy-issues-with-the-internets-most-popular-websites/

[25] Hirst, S. (2019). 8 Google Data Privacy Concerns You Should Be Worried About!. Retrieved from https://thevpn.guru/8-google-data-privacy-concerns/

[26] Google Technologies. (2019). Retrieved from https://policies.google.com/technologies

[27] Google Transparency Report. (2019). Retrieved from https://transparencyreport.google.com/user-data/overview

[28] Regulation (EU) 2016/679 of the European Parliament and of the Council 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation