Scientific Research Publishing

Journal of Information Security, 2017, 8, 56-74
http://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Web Search Query Privacy, an End-User Perspective

Kato Mivule

Department of Computer Science, Norfolk State University, Norfolk, USA
Email: kmivule@nsu.edu

How to cite this paper: Mivule, K. (2017) Web Search Query Privacy, an End-User Perspective. *Journal of Information Security*, **8**, 56-74.
http://dx.doi.org/10.4236/jis.2017.81005

Received: December 28, 2016
Accepted: January 14, 2017
Published: January 17, 2017

Abstract

While search engines have become vital tools for searching information on the Internet, privacy issues remain a growing concern due to the technological abilities of search engines to retain user search logs. Although such capabilities might provide enhanced personalized search results, the confidentiality of user intent remains uncertain. Even with web search query obfuscation techniques, another challenge remains, namely, reusing the same obfuscation methods is problematic, given that search engines have enormous computation and storage resources for query disambiguation. A number of web search query privacy procedures involve the cooperation of the search engine, a non-trusted entity in such cases, making query obfuscation even more challenging. In this study, we provide a review on how search engines work in regards to web search queries and user intent. Secondly, this study reviews material in a manner accessible to those outside computer science with the intent to introduce knowledge of web search engines to enable non-computer scientists to approach web search query privacy innovatively. As a contribution, we identify and highlight areas open for further investigative and innovative research in regards to end-user personalized web search privacy—that is methods that can be executed on the user side without third party involvement such as, search engines. The goal is to motivate future web search obfuscation heuristics that give users control over their personal search privacy.

Keywords

Web Queries, Web Search Privacy, User Profile Privacy, User Intent Privacy

1. Introduction

Search engines have become a useful part of a daily routine when it comes to searching for information on the Internet. However, the issue of privacy remains a major concern, due to the capability of search engines to retain user search logs. While search engine query log retention abilities might offer better personalized search results, user privacy is never guaranteed. Another challenge is that due to the enormous computation and storage power of search engines, query disambiguation keeps improving, making it problematic for users to reuse the same obfuscation techniques over time. Although research in web search query obfuscation has gained the attention of researchers [1] [2] [3] [4] [5], studies have noted that web search query confidentiality continues to be a difficult problem, mainly due to the monetization of search results by search engines [6] [7]. For instance, on the rationale for retaining user web search query logs, search engine companies offer the following rationale for doing so [8]: (i) Enhancing ranking algorithms, (ii) Query fine-tuning, (iii) Improving personalized query results, (iv) Combating fraud and abuse, (v) Enabling shared data for research, and (vi) Enabling shared data for marketing and other commercial purposes. It is interesting to note that each of the mentioned reasons for retaining user search query logs is a privacy concern. Even when organizations claim to privatize web search query logs, errors can still be made; as was the case with the 2006 AOL scandal in which a user was re-identified and traced to their geo-location after an anonymized set of web search query logs was published [9] [10].

Therefore, the user-based privatization techniques that do not require third party intermediaries are urgently needed as another layer of protection. Moreover, from a policy point of view, researchers have highlighted a number of relevant issues, important for gauging privacy guarantees when it comes to implementing web search query obfuscation methods. For example, Cooper (2008) noted that web search query obfuscation techniques could be judged using the following criteria [8]: (i) Effectiveness of the method to protect user privacy, (ii) Effectiveness of the procedure to conserve the usefulness of query results, and (iii) How effectively the user can have control to implement the privacy technique. As noted by Cooper (2008), the reasons given by search engines for retaining user web search query logs, are often wanting, with no user confidentiality guarantees. As a contribution, we identify and highlight areas open for further investigative and innovative research in regards to personalized web search privacy—that is methods that can be executed on the user-side without third party involvement such as, search engines. The goal is to motivate future web search obfuscation heuristics to give users control over their personal search privacy. The central question being asked by this study is if it is possible to generate web search query obfuscation methods that can be executed on the user-side without third party collaboration. For trusted privacy, users require techniques executed on the user side of the machine without involving untrusted third parties, such as search engine providers. This study reviews material in a manner accessible to those outside computer science with the intent to introduce knowledge of web search engines to enable non-computer scientists to approach web search query privacy innovatively. To reach readers outside computer science but interested in web search privacy, we have broken this article down into a review of web search engine mechanisms and suggested areas for further study. Therefore this study reviews web search engines with respect to web search queries and user intent privacy. While a number of cryptographic and anonymous web browsing techniques, such as Tor, have been suggested [11] [12] [13] [14], this article emphasizes privatized web search querying techniques that do not necessarily involve cryptographic and other third party anonymous web browsing methods. The rest of the paper is organized as follows. Section 2 reviews search engines and web search queries. Section 3 identifies areas for further research. Section 4 concludes the article.

2. The Web Search Engine

This section presents an overview on how search engines work; an essential foundation for formulating end-user personalized web search privacy techniques. A basic understanding of web search engine mechanisms is necessary, for example, for data privacy curators to formulate new web search query obfuscation methods with enhanced query result usability when compared with existing methods.

2.1. How Web Search Engines Work

A web search engine is a software tool used to search for information on various subjects on the Internet, and returns the most relevant search results to the user. A typical search engine works by providing the following functionality, as illustrated in Figure 1 [15] [16] [17]. It is important to note that each component of a search engine is a privacy concern as search engines maintain logs of user interactions with these components. While the explicit mechanics of how search engines work is beyond the scope of this paper, we endeavor to cover the web search query functionality of the search engine in greater detail since given that our concern is the privacy of search queries:

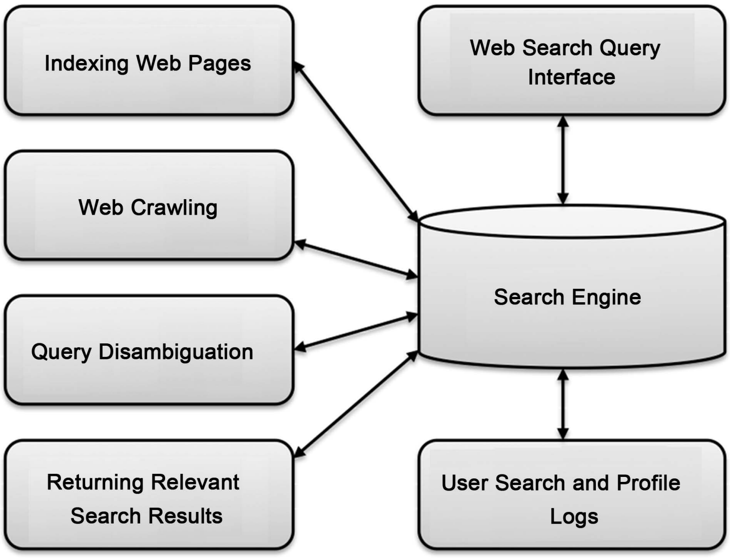

Indexing Web Pages
Web Search Query Interface
Web Crawling
Search Engine
Query Disambiguation
Returning Relevant Search Results
User Search and Profile Logs

Figure 1. Typical search engine functionalities.

- *Web crawling*—Web crawling involves a software agent that is given starting URLs and downloads every webpage by following each URL link. The URL of every downloaded webpage is stored and the document saved in a repository.
- *Indexing web pages*—Every downloaded webpage is indexed and stored in a repository. Each word on the downloaded page is stored, given a word identifier, sorted; and the webpage is given a document identifier.
- *Web search querying*—Search terms from the user are converted to word identifiers and the indexed documents are searched until a match occurs.
- *Relevant search results*—The search engine employs techniques and metrics, such as user profiling, user intent estimation, and Page Rank to return the most relevant search results to the user. A rank computation of the matching documents to the query is performed, and the *top-k* document results are returned; where k represents the number of documents returned.

2.2. Web Search Engine Data Structures

Web search engines include the following data structures for enhanced functionality, a key consideration for researchers when creating query obfuscation techniques to implement privacy [16] [18] [19]:

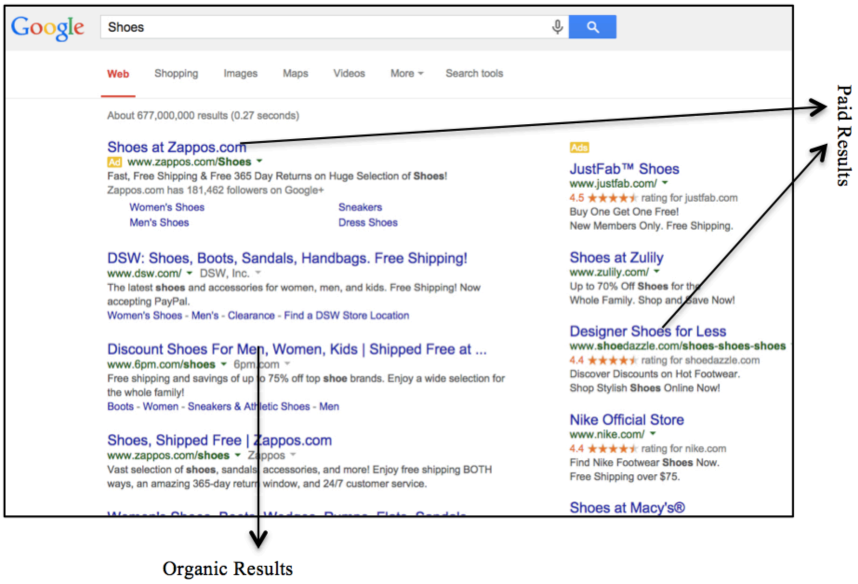
- *Repository*—The repository contains the full compressed HTML of every webpage, with its URL and a given document identifier.
- *Document Index*—The document index contains every webpage or document indexed sorted by document identifiers and a URL list containing the URLs.
- *Lexicon*—A lexicon is maintained with a list of every known word; Google's lexicon contained 14 million words by 1999.
- *Hit Lists*—Hit lists keep track of every occurrence of a word in a document.

2.3. Web Search Engine Navigation

Another area of concern in regards to end-user privacy is web search navigation. Search engines seek to profile users so as to correctly target advertisements and maximize revenue. When a user enters a search query, the search engine returns results that can be categorized in two of the following groups [20] [21]:

- *Paid search results*—These are targeted results based on user profiles, intent, and query search terms, generated and paid for by advertisers, as illustrated in **Figure 2**. User profiles in this context are behavioral profiles of the user generated from web browsing history and patterns over time, and thus a major privacy concern since they would collectively reveal a user's intent [22].
- *Organic search results*—These are general results returned by the search engine based on the user query, as shown in **Figure 2**.

  Organic search results can be further divided into these categories [20]:
- *Meta-search results*—These are a combination of search results collected from a group of search engines.
- *Grouped search results*—These are retrieved results after result clustering and classification.
- *Personalized search results*—These search results are generated using user search query logs, browsing history, user profiles, and usage records.



About 677,000,000 results (0.27 seconds)

Shoes at Zappos.com
Ad www.zappos.com/**Shoes**
Fast, Free Shipping & Free 365 Day Returns on Huge Selection of **Shoes**!
Zappos.com has 181,462 followers on Google+
Women's Shoes Sneakers
Men's Shoes Dress Shoes

DSW: Shoes, Boots, Sandals, Handbags. Free Shipping!
www.dsw.com/ DSW, Inc.
The latest **shoes** and accessories for women, men, and kids. Free Shipping! Now accepting PayPal.
Women's Shoes - Men's - Clearance - Find a DSW Store Location

Discount Shoes For Men, Women, Kids | Shipped Free at ...
www.6pm.com/**shoes** 6pm.com
Free shipping and savings of up to 75% off top **shoe** brands. Enjoy a wide selection for the whole family!
Boots - Women - Sneakers & Athletic Shoes - Men

Shoes, Shipped Free | Zappos.com
www.zappos.com/**shoes** Zappos
Vast selection of **shoes**, sandals, accessories, and more! Enjoy free shipping BOTH ways, an amazing 365-day return window, and 24/7 customer service.

Women's Shoes, Boots, Wedges, Pumps, Flats, Sandals

Ads

JustFab™ Shoes
www.justfab.com/
4.5 ★★★★★ rating for justfab.com
Buy One Get One Free!
New Members Only. Free Shipping.

Shoes at Zulily
www.zulily.com/
Up to 70% Off **Shoes** for the Whole Family. Shop and Save Now!

Designer Shoes for Less
www.**shoe**dazzle.com/**shoes-shoes-shoes**
4.4 ★★★★★ rating for shoedazzle.com
Discover Discounts on Hot Footwear.
Shop Stylish **Shoes** Online Now!

Nike Official Store
www.nike.com/
4.4 ★★★★★ rating for nike.com
Find Nike Footwear **Shoes** Now.
Free Shipping over $75.

Shoes at Macy's®

Figure 2. Paid search and organic search results.

- *Natural Language search results*—These search results involve question answering in a natural language.
- *Image search results*—These search results involve a return of image and multimedia to search queries.

2.4. Paid Search Engine Results Process

When a user executes a web search query, search engines provide advertising in the form of texts and images on the search results pages, as in Figure 2. Paid search results are often driven by user query inputs. For example, if a user searches for "Toyota", paid search results would include ads about where to buy a Toyota. Paid search results generally depend on the following [23]:

- *The advertiser*—These are entities that provide the advertisement source; the ads are usually temporal and thematic in nature, for example, a Samsung ad during the Christmas holidays would include Christmas sales.
- *The search engine*—The search engine acts as a middleman, targeting advertiser ads to specific audiences and users, based on data collected from user profiles, query, and browsing history. If a user's search history includes a large amount of queries on cars, then the search engine would target "Honda" ads when the same user searches for "Honda".
- *Users*—Users are visitors who utilize the search engine to query for information; search engines largely seek to understand user intent so as to correctly target advertisements to the right audience.

Paid search results could be used as one aspect in monitoring the effect of implemented web search query privacy, since advertisements are always targeted to specific users based on their profiles and search queries.

2.5. Ranking Factors

Search engines return user query results based on relevance. Relevance is affected by the search results' importance. Factors that might influence search results' importance include the following [20] [24]:

- *The webpage*—how frequently a word is used on a specific page.
- *The website*—the worth and authority of the website—e.g. New York Times vs. a blogger page.
- *Click-through data*—the number of clicks on a page—higher is better.
- *Social reference*—how frequently a website is mentioned on social media.
- *Geographical Location*—search engines tend to return local answers for local questions.

Factors, such as click-through-data and geographical location, are used in the generation of user-profiles by search engines, and therefore need to be given consideration when formulating end-user web search obfuscation methods.

2.6. User Profile Generation

Another essential aspect of search engines that requires privacy research attention is the generation of user profiles, mainly used for targeting advertisements and improving relevant search results. Search engines generate user profiles by employing these two methods [25] [26]:

- *Click logging*—the search engine tracks each URL from the search query results based on user browsing clicks. Click logging is used to build a user web search profile and to infer user intent by the search engine to reduce irrelevant results. The method works best when a user repeats the same query a number of times or refines the query.
- *Profile generation*—search engines employ user personal information, query history, browsing history, click-through history, bookmarks, download history, etc., in the generation of a user web profile.

2.7. Web Search Queries

The purpose of search engines is to provide functionality for users to query for information on the Internet. Insight into the workings of web search queries is fundamental for developing effective obfuscation techniques. Web search queries are words, phrases, or descriptions that a user inputs in a search engine that are matched against documents indexed and stored by search engines, to return results relevant to users [27]. Web search queries can be in the form of hyperlinks, but differ from basic database search queries in that they do not use strict syntax rules, as in SQL [28]. There are three main web search queries categories [29] [30] [31]:

- *Informational queries*—web search queries concerned with larger general topics, and return large related result numbers, e.g. "cars" and "travel".
- *Navigational queries*—these queries are concerned with finding a single website or webpage of particular inquiry, *i.e.* "Twitter" and "Yahoo Movies".
- *Transactional queries*—these query types indicate the user seeks to make an online action like an item purchase, downloading music, or viewing a movie; e.g., "buy a Toyota 2014".

2.8. Ranking Factors

The three main categories of web search queries could further be broken down into the following types of queries:

- *Boolean queries*: Boolean queries involve the use of logical operators in the query construction. The two common Boolean logical operators include the AND logical operator, which returns restrictive and exclusive search results, and the inclusive OR operator. Boolean search queries have been found to be non-intuitive and difficult for users to employ [27].
- *Faceted queries*—Another category of web search queries consists of faceted queries, in which the query is divided into subjects in combination with Boolean operators, with the goal of the user viewing all the documents. An example would be, {Car AND Fuel}, {Toyota OR Honda}, {Camry and Civic} [27].
- *Concept-based search query*—Concept-based search queries employ semantic concepts and ideas rather than keywords in the composition of queries to retrieve documents in the same concept area [31].
- *Single and Multi word sense queries*—Single word sense queries are composed of a word but with wide-ranging contextual meanings. Multi word sense queries are made up of multiple single-word sense queries, providing better context to meaning than a single word query [33].
- *Keyword-based queries*—Keyword queries are composed of words or terms that tend to be short, ambiguous, and focused on a particular entity or subject. Examples include, {Toyota, Honda, Sales}. However, keyword-based queries can have several acceptable inferences, and various interpretations, generating a large set indicating what that particular query could mean [34] [35].
- *Single and multi word queries*—Single word queries are composed of only one term, while multiple have many search terms. Single and multi word search queries might not be entity specific as keyword-based queries [36].
- *Context queries*—these are queries where the search is done in context with information, such as the user's profile, search history, browsing habits, query features, user background, user interests, etc., used to disambiguate the query and return results relevant to the user [37].
- *Natural language queries*—natural language queries are composed of search terms in the form of real questions in the user's natural language, without the use of query syntax and special formats. *i.e.* a user could query, "what is the circumference of the Earth?" [27] [38].

Web search queries tend to be brief, vague, consist of subtopics, and generalized into two major groups—faceted and ambiguous queries [30]:

- *Faceted queries*: Faceted queries can be composed of subtopics, but are non-ambiguous, clear, and return precise and relevant results to the user.
- *Ambiguous queries*: Queries usually have more than one meaning, and so the search engine returns results that might not be relevant to the user.

A categorization of ambiguous queries is needed to obfuscate web search queries, as noted by Song's taxonomy (2007) [39] [40] [41]:

- *Category A*: *Ambiguous Queries*—queries consisting of one or more search terms where each term has multiple meanings. Ambiguity results from the multiple meanings, *i.e.* "bank" could be a financial institution or riverbank.

- *Category B*: *Broad Queries*—these query types are composed of different subtopics a user might search for. Examples include, "race cars". Ambiguity is caused by the vagueness of the topic and subtopics. For example, a user might have differing meanings for keywords, "race" and "cars".
- *Category C*: *Clear Queries*—these are queries composed of keywords with a very narrow and specific meaning. Examples include, "Harvard University". Clear queries return numerous search query results with a higher degree of quality than the ambiguous and broad queries.

Under the natural language processing (NLP) queries category, Wu *et al.* (2014) further categorized question retrieval queries into two major groups [42]:

- *Short queries*: these are NLP type queries in which the questions being asked by the user are very short, unclear, and ambiguous, thus making it difficult to correctly pinpoint user intent.
- *Long queries*: in this group of NLP queries, the question asked by the user is long and complete, often resulting in precise and relevant search results.

2.9. Semantic Search

An important aspect of web search queries lies in semantic search. Semantic search is concerned with meaning, the understanding of an expression, insinuation, and/or inference, highlighting the relationship between similar keywords and phrases in a web search query [43]. Search engine organizations spend considerable effort in employing semantic search techniques to efficiently pinpoint user intent, return relevant results, and better target advertisements. Thus, for effective web search privacy and query obfuscation, proposed frameworks must take into consideration query semantics to enhance user privacy. The following are some of the relevant semantic search characteristics:

- *Disambiguation*: Semantic search goes through disambiguation, to remove any ambiguity and multiple word meanings, to return the most probable search term meaning [44] [45].
- *Generation of relevant results*: To generate relevant web search results, semantic search systems take into consideration the context of the search, location, intent, word alternatives, word substitutes, and generalized and specific word concept equivalents [46].
- *Natural Language Processing*: In semantic search, Natural Language Processing linguistic components, such as, homonymy and synonymy, are employed in efforts to better understand the meaning of a user query, to accurately predict user intent, and return relevant query search results [44].

2.10. Word Sense Disambiguation

Word sense disambiguation is a computational process of finding the contextual meaning of words, a problem that researchers have noted to be intractable, as are other difficult problems in artificial intelligence [47]. Word sense disambiguation requires external repositories of knowledge, such as thesaurus, ontologies, and corpora to get the right context and sense of words [47]:

- *Thesaurus*—a text repository that offers synonyms, which are similar word meanings, and antonyms, which are opposite word meanings.
- *Ontologies and Lexicons*—these are descriptions of concepts of particular subjects of interests, taxonomies, and semantic relations, such as WordNet.
- *Corpora*—this is a collection of texts for studying language representations.

External knowledge repositories are crucial to effectively understand user intent and enhance personalized search results. However, the use of word sense disambiguation repositories, such as thesaurus, ontologies, and corpora, is also vital for web search query obfuscation during the query formation and reformation process. The challenge is to formulate queries so that user intent is protected, but with less disambiguation to retrieve relevant search results and thus improve usability.

2.11. The Word Net Database

WordNet is a commonly used external repository of knowledge, employed in the process of query refinement and keyword reformation to get the appropriate word meanings. The following is a description of the possible semantic relations that could be derived when using WordNet, with applications in query disambiguation, and, more importantly, query obfuscation insofar as this study is concerned [48]:

- *Synonym*—these are words with same meaning. For example, *search* can be replaced with *investigate*.
- *Hyponym*—the first word is an exact occurrence of the second word. For example, *crimson* and *red*.
- *Hypernym*—the second word is an exact occurrence of the first word. For example, *Tablet* and *iPad*.
- *Meronym*—the first word is a component part of the subsequent word. For example, *foot* and *leg*.
- *Holonym*—opposite to meronyms, the second word is a component part of the first word. For example, *keyboard* and *keys*.

2.12. Web Search Query Disambiguation

Web search query disambiguation involves the reformation and refinement of query search terms to remove ambiguity, better predict user intent, and return the most relevant search results to the user. Web search query disambiguation involves the following techniques [43]:

- *Manual query modification*—the user makes modification to the query by adding or removing search terms.
- *Query expansion modification*—search terms are added to the original query with the use of semantics.
- *Query trimming modification*—some query search terms are removed to improve query results.

- *Conjunction and disjunction modification*—the conjunction (AND) is used in the query to combine search terms and return exclusively unambiguous results; the disjunction (OR) is used to return inclusively generalized results.
- *Substitution modification*—query search terms are replaced with similar search terms using semantic techniques.
- *Graph-based modification*—graph theory techniques are used so documents are viewed as nodes in the graph, and query search terms are employed to return semantically relevant and related documents in the graph.

2.13. Web Search Query Reformation

As mentioned earlier, one of the most essential search engine tasks is to understand user intent and yield search results that are most pertinent to the user. Therefore, understanding query reformation provides a facet into how search engines "think", in regards to capturing user intent. To achieve this, search engines employ web search query improvement techniques, in which user queries are refined and modified to accurately capture user intent. However, search engines also take advantage of and store web search query reformations performed by the user, to correct errors, typos, and for modifications of the query for personalized results. While search engines might return better and more specific search results to the user, web search query reformation could be used as an attack against web search query obfuscation [49] [50] [51] [52]. When a user modifies an obfuscated query, it might be possible for the search engine to deduce the real query from dummy queries in such instances. Therefore, it would be essential to understand and consider query reformation techniques, and design suitable web query obfuscation methods. A taxonomy of web search query reformation was compiled by Huang and Efthimiadis (2009), on how search engines could use such categorization to improve web search results [53]. Huang and Efthimiadis (2009), outlined 13 categorizations of query reformation that a search engine could use to detect, estimate user intent, and offer better search results [53]: *Word rearranging*—words in the initial query are reordered but unchanged in the reformulated query. *Whitespace and punctuation*—whitespaces and punctuations are changed or removed in the reformulated query. *Deleted words*—some words are deleted from the initial query but the same words are kept in the reformed query. *Supplemented words*—extra words are added to the initial query. *URL removal*—the URL is removed from the query. *Stemming*–word stems are altered in a query. For example, *jumping over boxes* to *jump over a box*. *Acronym formation*—the query words are transformed into an acronym. For example, the *United States of America* is altered to *USA*. *Acronym expansion*—the modified query will have an expansion of the acronym that appeared in the initial query. *Substring*—the subsequent query is a rigid prefix or suffix of the initial query. For example, *food restaurant* is changed to *food rest*. *Superstring*—the subsequent query contains the initial query as a prefix or suffix. For example, *food rest* becomes *food restaurants*. *Abbreviation*—matching words in the initial query are a prefix of every word in the subsequent query. For example, *higher sec* is changed to *high security*. *Word replacement*—words in the initial query are replaced with semantically related words in the subsequent query. *Spelling correction*—the Levenshtein distance algorithm is used to predict the spelling correction a user would do in the subsequent query, by counting the number of characters between the two queries. If the Levenshtein distance is equal or less than two, the swapping takes place. For example, *corection* is replaced with *correction*. In so far as search query obfuscation is concerned, Huang and Efthimiadis (2009) observed that there exists classes of query reformations that are difficult for classifiers to detect [53]. These involve: *Queries with semantic rephrasing*—the more complex a query is rephrased, the harder it is to classify. *Multi-reformations*—reformation techniques used to modify the query. Classifying multi-reformation queries is difficult because they lack a commutative property. Subsequent queries can yield different results from the initial query [53].

2.14. User Intent Based Query Temporality

Generally search engines track users by generating a behavioral user profile based on the user's search history, using techniques such as, cookie tracking, browser preferences, IP address geo-location, and URL clicks [22] [54] [55]. However, one way to understand user intent in web search queries is to study time-related search words and search phrases in a submitted query. Since user intent can be deciphered by looking at the temporality of a web search query, temporality must be considered for better search query privacy. The intent of a query, as related to time, can be categorized as follows [56] [57] [58]:

- *Implicit temporal intent*—in which the user specifies no time data in the query phrase. For example, the "Olympics".
- *Explicit temporal intent*—whereby the user makes specific time reference in a query. For example, the "2016 Olympics".

Queries with implicit temporal intent can be further broken down according to the following classifications [56]:

- *Atemporal*–queries that lack time information, e.g., "Olympic games";
- *Temporal unambiguous*—queries with specific time related information. For example, the "2012 Olympics";
- *Temporal ambiguous*—queries with unclear multiple time requests, *i.e.*, a query, "Isaac Newton", may return his birthday and discovery times.

2.15. Search Query Classification and Paid Ads

Another aspect important for web search query obfuscation is query classification. Web search query classification is an ongoing research challenge that involves the grouping of user queries into specific categories so as to better predict user intent, retrieve the most relevant webpages for the user, and direct web advertisements to the appropriate audiences [23] [59]-[65]. It is estimated that the most popular queries only involve between 2.4 and 2.7 words, making it difficult to disambiguate and pinpoint user intent, due to the small amount of information contained in the queries [23]. The main goal of web search query classification is to disambiguate queries, pinpoint user intent, and accurately direct paid search results to users [23]. However, Gabrilovich *et al.* observed that the problem of query classification remains intractable due to the short composition of queries. Moreover, Gabrilovich et al., pointed out that since search engines catalog huge quantities of information and, in so doing, become storehouses of knowledge, it therefore makes sense to use web query search results to get an understanding which can lead to query interpretation [23].

3. Areas for Further Investigative Study in Web Search Privacy

This section identifies areas for further study and investigative research. While some ongoing research covers potions of the proposed study areas listed below, we believe that the identified areas need further investigation, given the intractability and complexity involved in query ambiguity and disambiguation, and the trade-offs required to find the right balance between privacy and usability. The context here is that research could focus on non-cryptographic solutions that do not require the use of third party applications. Under the web search engines and web search query mechanism, we identified the following privacy challenges, as summarized in **Table 1**.

Table 1. A summary of web search engine and query privacy challenges.

Web Search Engine and Query Functionalities	Privacy and Usability Challenges
1. *Web search engine navigation privacy*	Holistic privacy and usability approach to web search navigation using query formation and reformation plus result selection and navigation.
2. *Paid and organic search results*	Employing paid and organic search results as a measure/indicator of web search query obfuscation.
3. *Precision and recall*	Using precision, recall, and Page Rank to measure web search query obfuscation and usability.
4. *User profile generation*	Generation of obfuscated user profile with an acceptable level of search result usability.
5. *Web search queries*	Privacy and usability approaches to informational, transactional, and navigational queries.
6. *Types of search queries*	Privacy and usability methods, including hybrids—e.g. Boolean, concept and keyword queries.
7. *Semantic search and word sense disambiguation*	Privacy and usability approaches to semantic search using statistical, natural language processing, graph theory, and machine learning models.
8. *Web search query reformation*	How to achieve privacy and usability during web search query reformation. How can a user correct their query without revealing their intent?
9. *Temporal queries*	How can an acceptable level of usability be achieved with obfuscated temporal queries?
10. *Query classification*	How to obfuscate user intent with query classification. Methods to obfuscate search queries using both semantics and different classifications.

- *Web search engine navigation privacy*: Further study is needed into holistic privacy and usability techniques that take the web search engine navigation process into account, including, (i) query formulation, (ii) result selection, (iii) result navigation, (iv) and query reformulation.
- *Employing paid and organic search results to enhance privacy*: Studies could include how to use paid and organic search results as a measure of obfuscation effectiveness. For instance, if paid results accurately mirror the real query instead of the dummy query then we can infer that the search engine has deciphered and separated dummy queries from real queries.
- *Precision and recall*: Questions, such as how to improve the precision and recall of noisy web search queries with respect to user expectations and results, warrant further studies. A fruitful approach could involve the application of new or innovative statistical models and empirical application results.
- *Obfuscated user profile generation*: Investigations could be made into whether it is possible to generate pseudo-user profiles while offering more personalized search results. This line of research has the potential to offer privacy with improved usability for obfuscated user profiles.
- *Web search query obfuscation*: This remains an open area for further investigation. One study area asks how to generate privatized web search informational, navigational, and transactional queries with better usability. Transactional queries would be of most interest, especially with applications to transactional databases, such as Amazon and social networks.
- *Types of search queries*: Another open area ripe for study and investigation is the obfuscation of different types of queries. For example, studies could include obfuscating Boolean queries, privatizing keyword queries, a hybrid of concept such as keyword queries for obfuscation, and, most challenging, how to obfuscate natural language queries when real questions are asked.
- *Semantic search and word sense disambiguation*: While some significant work has been done, it remains open to further investigation due to the challenges involved in query disambiguation. Questions include how to find a balance, with trade-offs, between ambiguity and disambiguation needs during the query obfuscation process. Other investigations could focus on using mathematical and statistical modeling, natural language processing, graph-theory, and machine learning techniques for usability-aware web search query obfuscation.
- *Privacy in web search query reformation*: There is not much research in privacy in web search query reformation. Search engines spend considerable effort disambiguating queries, and data mining user intent, with every query reformation. More studies are needed on how to achieve privacy during web search query reformation. Questions could include how web search query reformation methods could be used in conjunction with privacy methods to implement obfuscation techniques pre or post reformation. These techniques could be broken down into off-line vs. real time obfuscated query reformation.

- *Privacy of user intent in temporal queries*: Temporal queries are vulnerable to information leakage due to date and time information in the query search terms. However, not much work has been done with respect to implementing privacy and obfuscation techniques for temporal queries. This is another area for investigation and study, with consideration to the usability of obfuscated temporal queries. If a user intended to search for "Honda 2015", in the original query, how would usability be attained in an obfuscated "Honda 2014" query?
- *Web* search *query classification*: Another area of study that has captured the interest of researchers, and is worth further investigation, is the classification of web search queries with the goal of capturing user intent. However, the classification of web search queries remains a challenge when privacy is considered. For example, is there a way to obfuscate queries using both semantics and different query classifications?

4. Conclusion

In this survey, we presented an overview of web search querying from the privacy perspective and identified areas that need further investigation insofar as web search query privacy is concerned. Covering all areas of web search querying is beyond this study's scope, but an effort was made to highlight key essentials to the formulating end-user search privacy techniques. An example of an area not covered in depth is the study of mathematical and statistical models for web search query and related obfuscation techniques, a subject left for further study. A second goal for this study was to review the material in a manner accessible to those outside computer science. The intent was to introduce knowledge of web search queries and search engines to enable non-computer scientists to approach web search query privacy innovatively. While there has been considerable research interest in web search query obfuscation, web search privacy remains a challenge with no proposed generalizable solution. Future work will include the study of mathematical and statistical models for web search query and related obfuscation techniques as we identified in the area that need web search query privacy innovation. Solutions tailored to specific domains might be more appropriate. For example, a specific search query obfuscation method for healthcare systems might not work well when making queries into social media postings. Additionally, addressing challenges of web search query privacy and usability in terms of human computer interaction will have to be considered for future work. For instance, one question could focus on how additional delay in processing web search query results after obfuscation could affect people's willingness to make the tradeoff between privacy and usability. Finally, researchers need to consider the computation and storage power of intelligent search engines. Search engines store search query logs and over time with ever-improving disambiguation and semantic techniques. It is believable that search engines can decipher user intent and separate dummy queries from real queries in many circumstances. Privacy researchers have to have techniques that can operate in the context of a smart search engine–one that combines artificial intelligence, natural language processing, high performance computing, etc. to decipher user intent. Research into obfuscation techniques that take into consideration such search engine dynamics are needed to further advance the web search query privacy domain.

Acknowledgements

We would like to acknowledge the Department of Computer Science at Norfolk State University for making this work possible.

References

[1] Zheleva, E.G. (2011) Privacy in Social Networks: A Survey. In: *Social Network Data Analytics*, 277-306.

[2] Gotz, M., Machanavajjhala, A., Wang, G., Xiao, X. and Gehrke, J. (2012) Publishing Search Logs—A Comparative Study of Privacy Guarantees. *IEEE Transactions on Knowledge and Data Engineering*, **24**, 520-532. https://doi.org/10.1109/TKDE.2011.26

[3] Chen, T., Boreli, R., Kaafar, M.A. and Friedman, A. (2014) On the Effectiveness of Obfuscation Techniques in Online Social Networks. In: *Privacy Enhancing Technologies*, Vol. 8555 LNCS, 42-62.

[4] Ruiz-Martínez, A. (2012) A Survey on Solutions and Main Free Tools for Privacy Enhancing Web Communications. *Journal of Network and Computer Applications*, **35**, 1473-1492. https://doi.org/10.1016/j.jnca.2012.02.011

[5] Toch, E., Wang, Y. and Cranor, L.F. (2012) Personalization and Privacy: A Survey of Privacy Risks and Remedies in Personalization-Based Systems. *User Modeling and User-Adapted Interaction*, **22**, 203-220. https://doi.org/10.1007/s11257-011-9110-z

[6] Pang, H., Xiao, X. and Shen, J. (2012) Obfuscating the Topical Intention in Enterprise Text Search. *IEEE 28th International Conference on Data Engineering* (*ICDE*), 1-5 April 2012, 1168-1179. https://doi.org/10.1109/icde.2012.43

[7] Hillard, D., Schroedl, S., Manavoglu, E., Raghavan, H. and Leggetter, C. (2010) Improving ad Relevance in Sponsored Search. In: *Proceedings of the third ACM International Conference on Web Search and Data Mining—WSDM'*10, ACM, New York, 361-370. https://doi.org/10.1145/1718487.1718532

[8] Cooper, A. (2008) A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective. *ACM Transactions on the Web*, **2**, 1-27. https://doi.org/10.1145/1409220.1409222

[9] Arrington, M. (2006) AOL Proudly Releases Massive Amounts of Private Data. Techcrunch.com.

[10] Barbaro, M. and Zeller Jr., T. (2006) A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*, p. C4.

[11] Gao, X., Yang, Y., Fu, H., Lindqvist, J. and Wang, Y. (2014) Private Browsing: An Inquiry on Usability and Privacy Protection. *Proceedings of the* 13*th Workshop on Privacy in the Electronic Society*, Scottsdale, 3 November 2014, 97-106. https://doi.org/10.1145/2665943.2665953

[12] Danezis, G. and Diaz, C. (2008) A Survey of Anonymous Communication Channels.

[13] Reed, M.G., Syverson, P.F. and Goldschlag, D.M. (1998) Anonymous Connections and Onion Routing. *IEEE Journal on Selected Areas in Communications*, **16**, 482-494. https://doi.org/10.1109/49.668972

[14] Ren, J. and Wu, J. (2010) Survey on Anonymous Communications in Computer Networks. *Computer Communications*, **33**, 420-431. https://doi.org/10.1016/j.comcom.2009.11.009

[15] Gordon, M. and Pathak, P. (1999) Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. *Information Processing and Management*, **35**, 141-180.

[16] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, **30**, 107-117.

[17] Ozcan, R., Altingovde, I.S., Cambazoglu, B.B., Junqueira, F.P. and Ulusoy, Ö. (2011) A Five-Level Static Cache Architecture for Web Search Engines. *Information Processing & Management*, **48**, 828-840.

[18] Chakrabarti, S., Berg, M. and Dom, B. (1999) Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. *Computer Networks*, **31**, 1623-1640. https://doi.org/10.1016/S1389-1286(99)00052-3

[19] Barroso, L.A., Dean, J. and Holzle, U. (2003) Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*, **23**, 22-28.

[20] Levene, M. (2010) An Introduction to Search Engines and Web Navigation. Wiley, Hoboken. https://doi.org/10.1002/9780470874233

[21] Jansen, B.J. and Molina, P.R. (2006) The Effectiveness of Web Search Engines for Retrieving Relevant Ecommerce Links. *Information Processing & Management*, **42**, 1075-1098. https://doi.org/10.1016/j.ipm.2005.09.003

[22] Chen, Y., Pavlov, D., Canny, J.F. and Ave, H. (2009) Large-Scale Behavioral Targeting Categories and Subject Descriptors. *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, 28 June-1 July 2009, 209-218. https://doi.org/10.1145/1557019.1557048

[23] Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L. and Zhang, T. (2009) Classifying Search Queries Using the Web as a Source of Knowledge. *ACM Transactions on the Web*, **3**, Article No. 5. https://doi.org/10.1145/1513876.1513877

[24] Nunney, M. (2012) SEO Made Simple, Wordtracker's Free SEO Guide.

[25] Speretta, M. and Gauch, S. (2005) Personalized Search Based on User Search Histories. *Proceedings of the* 2005 *IEEE*/*WIC*/*ACM International Conference on Web Intelligence*, Compiegne, 19-22 September 2005, 622-628. https://doi.org/10.1109/WI.2005.114

[26] Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004) Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. *Proceedings of the* 13*th Conference on World Wide Web*, New York, 17-20 May 2004, 675-684. https://doi.org/10.1145/988672.988764

[27] Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern Information Retrieval. ACM Press, New York.

[28] Lee, W.M. and Sanderson, M., (2010) Analyzing URL Queries. *Journal of the Association for Information Science and Technology*, **61**, 2300-2310. https://doi.org/10.1002/asi.21407

[29] Broder, A. (2002) A Taxonomy of Web Search. *ACM SIGIR Forum*, **36**, 3-10. https://doi.org/10.1145/792550.792552

[30] Ullah, M.Z. and Aono, M. (2014) Query Subtopic Mining for Search Result Diversification. *IEEE International Conference of Advanced Informatics*: *Concept*, *Theory and Application* (*ICAICTA*), Vol. 1, Bandung, 20-21 August 2014, 309-314.

[31] Zamora, J., Mendoza, M. and Allende, H. (2014) Query Intent Detection Based on Query Log Mining. *Journal of Web Engineering*, **13**, 24-52.

[32] Egozi, O., Markovitch, S. and Gabrilovich, E. (2008) Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, **29**, Article No. 8.

[33] De Luca, E.W. and Scheel, C. (2013) Disambiguate Yourself. *Translation*: *Computation*, *Corpora*, *Cognition*, **3**, 75-86.

[34] Demidova, E., Zhou, X., Oelze, I. and Nejdl, W. (2010) Evaluating Evidences for Keyword Query Disambiguation in Entity Centric Database Search. 21*th International Conference on Database and Expert Systems Applications*, Bilbao, 30 August-3 September 2010, 240-247. https://doi.org/10.1007/978-3-642-15251-1_19

[35] Pound, J. and Hudek, A.K. (2012) Interpreting Keyword Queries over Web Knowledge Bases. *Proceedings of the* 21*st ACM International Conference on Information and Knowledge Management*, Maui, 29 October-2 November 2012, 305-314.

[36] Liu, B. (2007) Web Data Mining. Springer-Verlag, Berlin Heidelberg.

[37] Croft, W.B. and Wei, X. (2005) Context-Based Topic Models for Query Modification.

[38] Hristidis, V. (2009) Natural Language Queries Information Discovery on Electronic Health Records. CRC Press, Boca Raton.

[39] Liu, Y., Song, R., Zhang, M., Dou, Z., Yamamoto, T., Kato, M., Ohshima, H. and Zhou, K. (2014) Overview of the NTCIR-11 IMine Task. *Proceedings of the* 11*th NTCIR Conference*, Vol. 14, Tokyo, 9-12 December 2014, 8-23.

[40] Luo, C., Liu, Y., Zhang, M. and Ma, S. (2014) Query Ambiguity Identification Based on User Behavior Information. In: Jaafar, A., *et al*., Eds., *Information Retrieval Technology*, Springer International Publishing, Basel, 36-47.

[41] Song, R., Luo, Z., Wen, J.-R., Yu, Y. and Hon, H.-W. (2007) Identifying Ambiguous Queries in Web Search. *Proceedings of the* 16*th ACM International Conference on World Wide Web*, Banff, 08-12 May 2007, 1169-1170. https://doi.org/10.1145/1242572.1242749

[42] Wu, H., Wu, W., Zhou, M., Chen, E., Duan, L. and Shum, H.-Y. (2014) Improving Search Relevance for Short Queries in Community Question Answering. *Proceedings of the* 7*th ACM International Conference on Web Search and Data Mining*, New York, 24-28 February 2014, 43-52. https://doi.org/10.1145/2556195.2556239

[43] Mangold, C. (2007) A Survey and Classification of Semantic Search Approaches. *International Journal of Metadata*, *Semantics and Ontologies*, **2**, 23. https://doi.org/10.1504/IJMSO.2007.015073

[44] Manning, C.D. and Schutze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge.

[45] Guha, R., McCool, R. and Miller, E. (2003) Semantic Search. *Proceedings of the* 12*th International Conference on World Wide Web*, Budapest, 20-24 May 2003, 700-709. https://doi.org/10.1145/775152.775250

[46] Bonino, D., Corno, F., Farinetti, L., Bosca, A., Torino, P. and Duca, C. (2004) Ontology Driven Semantic Search. *Transactions on Information Science and Applications*, **1**, 1597-1605.

[47] Navigli, R. (2009) Word Sense Disambiguation. *ACM Computing Surveys*, **41**, Article No. 10. https://doi.org/10.1145/1459352.1459355

[48] Miller, G.A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM*, **38**, 39-41. https://doi.org/10.1145/219717.219748

[49] Dang, V., Croft, W.B. and Croft, B. (2010) Query Reformulation Using Anchor Text. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, New York, 4-6 February 2010, 41-50. https://doi.org/10.1145/1718487.1718493

[50] Song, Y., Zhou, D. and He, L. (2012) Query Suggestion by Constructing Term-Transition Graphs. *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, Seattle, 8-12 February 2012, 353-362. https://doi.org/10.1145/2124295.2124339

[51] Gupta, M. and Bendersky, M. (2015) Information Retrieval with Verbose Queries. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, 9-13 August 2015, 1121-1124. https://doi.org/10.1145/2766462.2767877

[52] Bing, L., Lam, W., Wong, T.-L. and Jameel, S. (2015) Web Query Reformulation via Joint Modeling of Latent Topic Dependency and Term Context. *ACM Transactions on Information Systems*, **33**, Article No. 6. https://doi.org/10.1145/2699666

[53] Huang, J. and Efthimiadis, E.N. (2009) Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, 2-6 November 2009, 77-86. https://doi.org/10.1145/1645953.1645966

[54] Hannak, A., Sapiezynski, P., Kakhki, A.M., Krishnamurthy, B., Lazer, D., Mislove, A. and Wilson, C. (2013) Measuring Personalization of Web Search. *Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, 13-17 May 2013, 527-537. https://doi.org/10.1145/2488388.2488435

[55] Liu, F., Yu, C. and Meng, W. (2004) Personalized Web Search for Improving Retrieval Effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, **16**, 28-40. https://doi.org/10.1109/TKDE.2004.1264820

[56] Campos, R., Al, J. and Jorge, A.M. (2011) Using Web Snippets and Web Query-Logs to Measure Implicit Temporal Intents in Queries to Cite This Version : Using Web Snippets and Query-Logs to Measure Implicit Temporal Intents in Queries. *ACM SIGIR 2011 Workshop on Query Representation and Understanding*, New York.

[57] Lin, S., Jin, P., Zhao, X. and Yue, L. (2014) Exploiting Temporal Information in Web Search. *Expert Systems with Applications*, **41**, 331-341. https://doi.org/10.1016/j.eswa.2013.07.048

[58] Xu, Z., Liu, Y., Mei, L., Hu, C. and Chen, L. (2014) Generating Temporal Semantic Context of Concepts Using Web Search Engines. *Journal of Network and Computer Applications*, **43**, 42-55. https://doi.org/10.1016/j.jnca.2014.04.002

[59] Jansen, B.J., Booth, D.L. and Spink, A. (2007) Determining the User Intent of Web Search Engine Queries. *Proceedings of the 16th ACM International Conference on World Wide Web*, Banff, 8-12 May 2007, 1149-1150. https://doi.org/10.1145/1242572.1242739

[60] Ortiz-Cordova, A. and Jansen, B.J. (2012) Classifying Web Search Queries to Identify High Revenue Generating Customers. *Journal of the American Society for Information Science and Technology*, **63**, 1426-1441. https://doi.org/10.1002/asi.22640

[61] Rose, D.E. and Levinson, D. (2004) Understanding User Goals in Web Search. *Proceedings of the 13th ACM International Conference on World Wide Web*, New York, 17-20 May 2004, 13-19. https://doi.org/10.1145/988672.988675

[62] Beitzel, S.M., Jensen, E.C., Frieder, O., Grossman, D., Lewis, D.D., Chowdhury, A. and Kolcz, A. (2005) Automatic Web Query Classification Using Labeled and Unlabeled Training Data. *Proceedings of the* 28*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, 15-19 August 2005, 581-582. https://doi.org/10.1145/1076034.1076138

[63] Agrawal, R., Yu, X., King, I. and Zajac, R. (2011) Enrichment and Reductionism: Two Approaches for Web Query Classification. Proceedings of 18*th International Conference on Neural Information Processing*, Vol. 7064, Shanghai, 13-17 November 2011, 148-157.

[64] Broder, A.Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V. and Zhang, O. (2007) Robust Classification of Rare Queries Using Web Knowledge. *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, 23-27 July 2007, 231-238. https://doi.org/10.1145/1277741.1277783

[65] Cao, H., Hu, D.H., Shen, D., Jiang, D., Sun, J.-T., Chen, E. and Yang, Q. (2009) Context-Aware Query Classification. *The* 32*nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, 19-23 July 2009, 3-10.