ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

Imporve Quality and Privacy in Personalized Search

Khushal Rathod [1], Pratibha Chavan [2]

P.G. Student, Department of Information and Technology Engineering, R.M.D.Singhgad School of Engineering , Warje, Pune, Maharashtra, India[1]

Associate Professor, Department of Information and Technology Engineering, R.M.D.Singhgad School Engineering, Warje, Pune, Maharashtra, India[2]

ABSTRACT: As the size of the Internet continues to grow the users of search providers continually demand search results that are accurate to their needs. Personalized Search is one of the options available to users in order to sculpt search results returned to them based on their personal data provided to the search provider. This raises concerns of privacy issues however as users are typically uncomfortable revealing personal information to an often faceless service provider on the Internet. This paper aims to deal with the privacy issues surrounding personalized search and discusses ways that privacy can be enriched so that users can become more comfortable with the release of their personal data in order to receive more accurate search results.

KEYWORDS: Personalized, Encryption, Privacy.

I .INTRODUCTION

With the ever growing size of the Internet, finding the right information from the right sources will become increasingly difficult. Popular search engines such as Google and Yahoo! are always improving on their search algorithms and search engines, but this may only take them so far. With the amount of pure content alongside saturated amounts of advertisements and spam on the Internet, finding what the user is really looking for through standard search mechanisms can be quite difficult. A typical search engine cannot handle ambiguity to appease all possible users when a query term such as "Golf" is used for example. In one scenario, the user may be looking for information relating to the sport of "Golf" while another user may be inquiring about the automotive offering of a Volkswagen "Golf". This will lead to situations where users are both unhappy with their search results and the contextual ads, which search providers heavily rely on for revenue, will be ineffective. One way of handling this ambiguity is to supply more personal information about the client to the search provider. In this example, suppose we knew that the user was an avid golfer. Then results for "golf clubs" and the "sport of golf" would be opportune, while results for a Volkswagen Golf could be easily filtered out of the result set. This methodology of supplying personal information to a search provider to enrich a user's search is called Personalized Search and popular search engines have released products to provide this technology to their users [1] [2].

One of the inherent problems with personalized search is that users are often insecure about handing over otherwise private or personal information regarding themselves to a search provider. Intuitively, the more that a search provider knows about a specific user, the more accurate their search results can be tailored for them, but how are the users to trust that the information that the search provider maintains about them will not be mishandled, lost, or maliciously used? If users can trust their chosen search providers with their personal information, then the providers can use that to deliver more accurate results with more specifically tailored advertisements. Thus, it should be in the keen interest of all providers of personalized search mechanisms to enhance the user's privacy surrounding their personalized search services as much as realistically possible. In order to enhance this privacy, this paper will look at philosophies and methods to optimize the privacy that users are given when using a typical personalized search service.

DOI: 10.15680/IJIRSET.2014.0312019

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

When using a Personalized Search service such as the ones mentioned [1] [2], how is the user and the search provider to ensure the privacy and protection of a users identity and information that is supplied to the service? As mentioned before, if a user does not trust the search provider, than the user is simply not going to either a) divulge sufficient personal information in order to optimize his/her search results or b) will not use the personalized search feature at all. Also, users may have concerns regarding the security surrounding the storage of their data. After a user's personal information has been given away, the onus is on the search provider to ensure that the information remains private and does not fall into the hands of people or organizations with malicious intentions for that data.

II. LITERATURE SURVEY

There are two main methods to structure the personal information handling of a personalized search service. The first method keeps all personal user information on the search provider's side of the transaction, for the scope of this report this method will be called server-side personalized search. When a user of the personalized search service goes to perform a search, their submitted information will be retrieved from the appropriate sever(s) at the search provider (or some other company, organization or location depending on the information technology environment being applied) and then used to, if possible, sculpt the search results and the contextual advertising to suit that user. An alternative to this system, called client-side personalized search would have the user keep their own set of personal user information to send to the search provider. The user would then be required to send this data to the search provider at every instance they wish to perform a personalized search. Both of these methods will be analyzed and the privacy concerns of each method will be discussed.

A. Server-Side Personalized Search Strategy

When employing a server-side personalized search strategy, there are two main opportunities for the personal information submitted to the service to be compromised. The first vulnerable place is during the initial transaction when the user submits their set of personal information to the search provider. If this information is sent to the provider in simple plaintext, then the user's information can be easily intercepted via a packet sniffing mechanism and then used however the interceptor may see fit. The second opportunity for privacy to be lost occurs if a malicious security breach occurs on the servers that house the personal information for the users of the search provider. This breach could lead to the loss of any privacy that users believed they had with their personal information on the search provider's servers. One basic way of ensuring that users' personal data remains private, in lieu of the outlined security problems, is to encrypt the personal information while in transit between the client/server and while stored in the search provider's database. This method will prevent any personal user information from existing in a plaintext format which is intrinsically vulnerable. Methods to encrypt the personal information and transport it will be discussed later.

B. Client-Side Personalized Search Strategy

Client-side personalized search strategy avoids the privacy risk of storing personal user information on search providers' severs by letting the client maintain and be responsible for their own 'set' of personal information. With this information, the client transports it to the search provider whenever they perform a search. The search provider will then take the received personal information along with the search query and then perform a personalized search for the client. By allowing the user to maintain their own personal data it increases the privacy for the user and thus, the search provider will not have to store a copy of the data on their severs. This allows the search provider to avoid responsibility for the integrity and privacy of this data. This technique it does have a few limitations however. The first issue is that this process is bandwidth intensive. A server-side search strategy needs only to transmit the search query to the provider during each user session. Client-side strategy on the other hand will typically require, depending on how the personalized search service is engineered, the client to submit their set of personal information alongside each search query. Most often the personal information will be vastly larger than the simple 2-5 word search query that the user is submitting. This forces the search provider and the client to deal with a much larger workload of bandwidth then they would have to deal with otherwise. As for the actual privacy concern with this set up, by making the user submit their personal information alongside their search query at every instance increases the chance that the information could be intercepted, like mentioned before, by a packet sniffing mechanism. Unless the transmission was applying a basic security mechanism such as encryption (opposed to allowing the transmission to exist in plaintext), the user's personal

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

information for the search provider will be vulnerable more often than it would be if a server-side strategy was being applied.

II. ENCRYPTION TO SUPPORT PRIVACY

So far two methods have been discussed alongside their limitations that enhance the privacy surrounding the personal information that users submit to personalized search providers. In order to make either of these methods viable, encryption of user data must be introduced as the first line of defense to guard the privacy of users.

A. Securing Server-Side Personalized Search with Encryption

To secure server-side personalized search, two primary mechanisms need to be protected.

- Transportation of personal user information to the search provider.
- Storage of personal user information on the search provider's servers.

To enhance the privacy of the transport of personal user information a public-key cryptography message system can be implemented. Using public-key encryption, the following strategy can be used to secure the transport of private user information to the server.

i. User encrypts their personal information using a public key of the search provider.
ii. User sends ciphertext to the search provider.
iii. Search provider uses their private key to decrypt the user's personal information as needed.

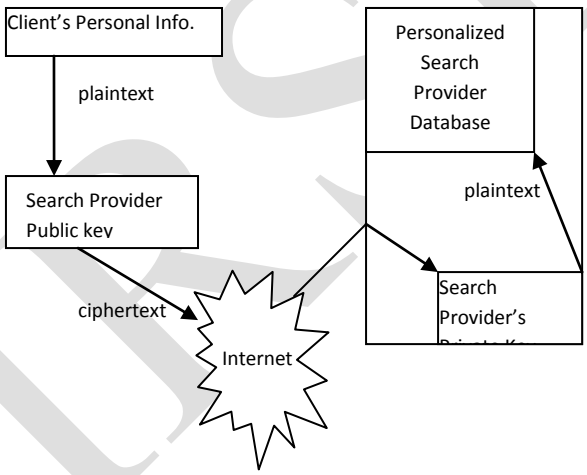

Figure 1 Personal Information Transmission to Search Provider

To handle privacy using encryption for storage of personal user information the following plan could be adopted.

i. Search provider encrypts all personal user information within their databases using their public key.
ii. When needed to perform a personalized search, the specific user's data is withdrawn from the database, decrypted with the search provider's private key and then fed into the program that performs the personalized search.
iii. The instance of that user's personal data that has been withdrawn and currently in plaintext will then be destroyed.

Having the personal information of users exist in plaintext for as little time as possible is the primary goal of this strategy to ensure user privacy. Providing that the search provider's private key can remain private, the provider should be able to maintain user privacy at all times. This system does not account for privacy breaches from within the actual search provider's organization however. An internal attacker may have access to the private key of the organization and thus, find a method of accessing the database and acquiring the personal information of their clients. This concern is out of the scope of this report however.

DOI: 10.15680/IJIRSET.2014.0312019

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

B. Securing Client-Side Personalized Search with Encryption

Securing client-side personalized search is similar to securing the transport phase of server-side personalized search. Each time the user performs a personalized search, the user's information for the search provider will have to be transported in the same fashion as outlined in Figure 1. The only difference here is that the search provider will return the user's queried results and then destroy the user information that was sent to them. As secure as this method may be, extra iterations of encryption and decryption will be necessary as the user is sending their encrypted personal information alongside each of their search queries. This limitation will increase the processor load on both the client machine and the server as they will continually have to encrypt and decrypt the transmissions respectively.

IV. EXPERIMENT OVERVIEW: SIMPLE PRIVACY OF GOOGLE WEB HISTORY

Google Web History is one of the server-side personalized search offerings that exist in the current market. It differs slightly in the discussed personalized search methodology because instead of relying on a simple user-defined profile to aid in the personalized search, Google Web History uses the user's previous search queries and the results that they clicked on in order to sculpt future queries. Thus, the more a user expands their Google Web History the theoretically more accurate Google's results for that user will be. Google's seemingly infinite recording of a willing user's search history can easily raise concerns of privacy and security surrounding search results. Fortunately Google has employed strong, but not bulletproof solutions to ensure user privacy.

1. Secure HTTP (HTTPS) login and authentication using AES-256 bit encryption. This is the first line of defense. Users must be logged into their respective Google account (identified typically by a *"@gmail.com" address) in order to view their web history.
2. Since users can remain logged into Google's services for extended periods of times, any request to view a logged in users web history after the immediate initial authentication will have the user required to reenter their password (again, using the security defined in the 1st point) in order to view their web history.

This seems like a sufficient way to ensure the privacy of user's personal information for the Google Web History service, but beyond the initial login, there ceases to exist any HTTP security or encryption within the actual Web History application. This means that any user with a simple packet sniffer can capture the web history of a user providing that they exist in the same Ethernet segment or using a wireless internet connection.

A.Experiment Process

To demonstrate the ease of this unauthorized access a simple experiment has been devised using my own Google Web History account and a tool called Wireshark, a freely available network analysis tool. To acquire the required packets the following configuration was used.

- Interface: *Broadcom 802.11b/g WLAN (Microsoft's Packet Scheduler) : \Device\NPF_{6FF15654-ECE8-4E1A-8B1B-E225218DB89B}*
- Promiscuous Mode: *true*
- Capture Filter: *tcp port http*

Once configured the packet capture tool was initiated and then a web browser directed to navigate to and log into Google Web History. Once a sample page of my own web history completed loading the packet capture tool was stopped. The next step is to analyze the captured packets and search for any information that might relate to my web search history recorded by Google's Web History product.

B .Experimental Result

During the experiment, Wireshark was able to acquire 25 packets and was able to compose seven TCP segments into one reassembled TCP HTTP result that contained the HTML page that listed of my Google Web History for September 23, 2007. To extract the list of web history from this file a simple 'Find' command was used to search for all instances of the token "http://www." within the file. This allowed the manual extraction of all URLs contained in the assembled HTTP page. From this page, a list of previous Google web searches and their results was composed.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

From this simple experiment, it should be recommended that Google "wraps" there entire Google Web History product in secure HTTP using AES-265 encryption (i.e. the same security that Google used for their login and authentication process for accessing the Google Web History service)

C. Future Experimental Considerations

To further assert the lack of sufficient privacy surrounding Google's Web History product larger scale and more in depth experiments would have to be performed. The experiment included herein only focused in a single client machine with the network analysis software running on that specific machine. A more through network analysis of Google Web History would include a sole computer dedicated to network packet capture while other 'done' clients were used on various different Google Accounts to access each user's respective web history. This process would give a more practical overview of the ease of acquiring private user information from the Google Web History tool. The expansion of this experiment would also call for a tool to be used to automatically find and extract valid URL's from the assembled or partially assembled HTTP results in order to vastly accelerate the identification of user's web history results from the set of captured packets.

V.CONCLUSION

When it comes to privacy of personalized search, users primarily need to be educated on the advantages and disadvantages of personalized search. Some users will be willing to sacrifice some of their online anonymity for better search while others may not find any advantage in this. Search providers also need to recognize the onus of responsibility in protecting their users' personal information at all costs. In this paper we analyzed two different methods for securing the privacy of personal user information in a personalized search environment. Although both methodologies had their advantages and limitations, server-side search personalization should be recommended over client-side personalization. Server-side implementations have the benefit of better performance and less processor usage than the demands of client-side search personalization. As the size and intensity of the Internet grows users may become more and more dependent on personalized search offerings. Hopefully in the future search providers will offer stronger privacy controls and agreements to ensure that all users of their service can completely trust the security and integrity of any personal and private information that they may submit to search providers.

REFERENCES

[1] Google Web History. http://www.google.com/psearch
[2] Yahoo!Search Builder http://builder.search.yahoo.com
[3] Xeuhua Shen , Bin tan ChenXiang Zhai , Department of Computer Science University of Illinois, "Privacy Protection in Personalized Search", ACM SIGIR Forum Vol.41 No.1, June 2007.
[4] Yabo Xu, Benyu Zhang, Zheng Chen, Ke Wang; Simon Fraser University and Microsoft Research Asia, "Privacy-Enhancing Personalized Web Search", International World Wide Web Conference archive proceedings of the 16th international conference on World Wide Web, 2007.
[5] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[6] K. Elissa, "Title of paper if known," unpublished.
[7] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[9] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.