SEARCH ENGINES: THE INVADER TO PRIVACY

Narendra Lahkar Sanjib K Deka

Abstract

The Search Engines has dominated the Internet by their popular social services. The web Search Engines has been regarded as the second most used web service after the popular Internet service, i.e., e-mail. Worldwide search markets shows that more than 750 million people - an estimated 95 percent of global Internet users conducted 61 billion cumulative searches in every month. In this paper it has been discussed the magnitude of Search Engine. It has been seen that the Search Engines keep track of all the searches that the user formulates. The paper discussed the possibility of invading privacy by the well-advanced Search Engines and possible leakage of personal information by these web search services. Some solutions have been discussed to cope up with these Search Engines threats.

Keywords : World Wide Web, Search Engines, Privacy Threats, Future Web Search

1. Introduction

An explosive growth has transformed the Internet from a specialized diminutive network not too long ago to its current ubiquitous position as a global fundamental tool in several disciplines, ranging from international business and e-commerce to education and everyday family life. Furthermore, simple observation yields that many of these sites comprise hundreds or even thousands of individual web pages. This sheer amount of accessible website content has brought about the undeniable need for fast, effective search engines. Otherwise, web users would find it practically impossible when facing the enormity of the World Wide Web to access the information that they require, as browsing randomly for information would clearly make no sense. In fact, an estimated 85% of all web users utilize services provided by Search Engines to locate desired information on the web, and search engine use is considered to be the second most popular Internet activity after e-mail use. But just how powerful and intrusive can they be, and how much of a threat against privacy? In principle, search engines listing rules, ranking rules, and advertising policies might shield users from some bad practices, and user's good judgment could protect them from others. The rise of paid search results brings additional complications. Prominent results often reflect solely a site's willingness to pay rather than its quality, relevance, or safety.

The technology employed by search engines is astonishing. For example, Google indexes over 8 billion web pages and 2 billion images. In March 2007 alone, it processed over 3.8 billion queries, equivalent to 1400 queries a second, making it a market leader with almost 54% of all searches conducted in that period. However, the same technology that is the foundation of search engines' success is increasingly giving rise to privacy concerns. Concerning privacy, search engines could be described as a 'data bottleneck'. Companies usually need to analyze lots of access statistics to create a profile for a single user. Yet when using a search engine the user himself submits condensed and accurate information on her topics of interest. This makes search engines an interesting target for profiling and advertising companies.

2. The Most Popular Search Engines and Emerging Trends of Search Engine Technology

It is interesting to mention which of the thousands of Search Engines are most favored by web users. The search engine field is fascinating to analyze, since many search engine sites use third-party searching technologies to provide the results, while in turn these external providers might run their own search engine sites as well. These facts, coupled with mergers and acquisitions amongst Search Engines, make up an intricate network of partnerships, collaborations and fierce competition.

Nielsen Online, a service of the Nielsen Company, delivers comprehensive, independent measurement and analysis of online audiences, advertising, video, consumer-generated media, has reported the following top 10 search providers in September 2007.

Top 10 Search Providers

Provider	**Searches(000)**	**YOYGrowth**	**Share ofSearches**
Google Search	3,994,158	41.3%	54.0%
Yahoo! Search	1,443,244	9.3%	19.5%
MSN/Windows Live Search	890,685	71.5%	12.0%
AOL Search	444,493	24.0%	6.0%
Ask.com Search	158,969	4.5%	2.2%
My Web Search	61,911	N/A	0.8%
Comcast Search	38,926	N/A	0.5%
BellSouth Search	35,740	30.3%	0.5%
SBC Yellow Pages Search	29,424	42.6%	0.4%
My Way Search	26,750	-78.4%	0.4%

Source: Nielsen Online, MegaView Search

Google is the most used search engine in the world, according to a report published by U.S. Internet research agency comScore. Its first comprehensive findings of worldwide search markets shows that in August 2007, more than 750 million people - an estimated 95 percent of global Internet users - conducted 61 billion cumulative searches. Google dominated search behavior; 37.1 billion searches were conducted from its websites. While majority came from its California based search engine, five billion were done on its video-sharing website YouTube. Yahoo was the second most used search engine with 8.5 billion queries. Chinese search engine Baidu ranked third handling 3.2 billion queries followed by Microsoft's MSN and Live Search Services with 2.2 billion. South Korean search firm Naver, owned by NHN Corporation, closed in at number five handling 2 billion queries. The findings also show that most search activity happens in the Asia-Pacific region, which has large markets like China, Japan and India where 258 million unique searchers conducted 20.3 billion searches.

Both the nature and scope of information on the Internet and available knowledge are evolving quickly. Future search services will no longer be restricted to conventional computing platforms. Engineers have already integrated automotive mobile data communications, known as Telematics, and it is likely they will also embed search capabilities into entertainment equipment such as game stations, televisions and high-end stereo systems. Also, with cameras becoming an integral part of mobile phones, Internet service providers are looking for new technologies in search engines to promote use of photo images. Thus, search engine activity has become a primary activity on the World Wide Web, where billions of documents are distilled down to approximately twenty to thirty links, depending upon the nature and type of search query. The popularity of the Internet is that it offers a variety of content not available in any other medium.

3. How Search Engines may Track the Users and their Privacy

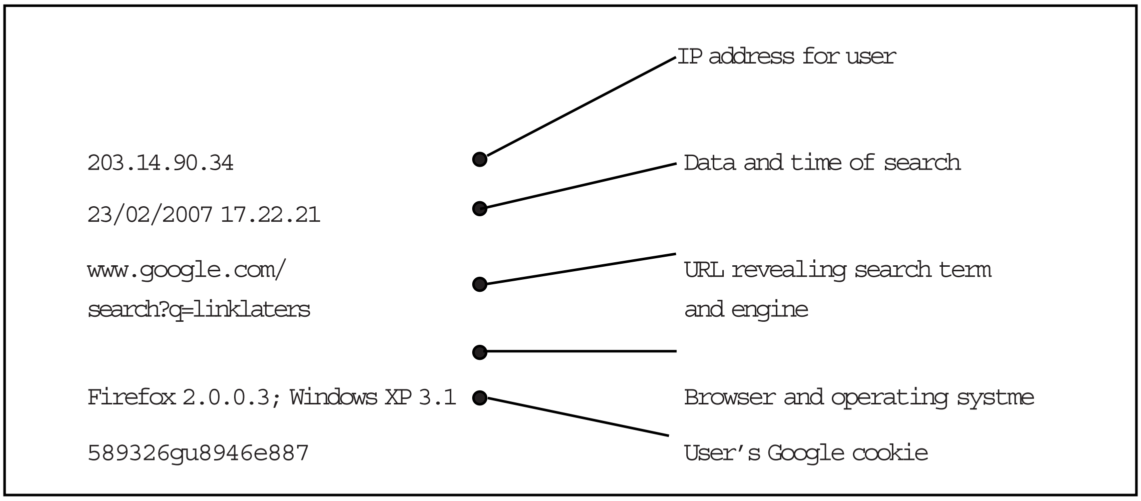
People often view search engines as benign blank boxes to which they can pose any question they want and not suffer the consequences. Search engines large and small typically keep logs of users search terms, with some search engines going further and matching those terms to users computer address, name, and other items, depending on how much information they have shared with the search engine. When a web surfer first visits a Search Engine, it will most likely log the IP address of the computer being used, the date and time of the access, and probably the browser configuration. If available, referrer information may also be logged, for if the user arrived at the Search Engine page by clicking on a link in some other web page, then this referrer data will contain the URL of this previously visited page. This analysis can yield important information for them, such as the sites that better feed traffic to their sites and the approximate regional location of the visitors.

However, the major difference between regular websites and Search Engines is that the latter have the possibility to keep track of all the searches that the user formulates during a visit to the site. All of the entered keywords can be traced, and even the links in the results page that were ultimately accessed by the user can be logged. This type of information collection comes off as very natural since Search Engines are merely keeping track of the service they are offering to the users, but it is one of the first practices carried out by Search Engine sites which is a matter of privacy concerned. Cookies appear simple enough on the surface. They are actually nothing more than small text files used to keep some information on the client computer. The main purpose of Search Engines when placing persistent cookies (which remain stored in the user's hard drive until erased or expired) is not to trace the search habits of their users per se, but rather to have a way of individualizing the preferences of those accessing the site, in order to provide a more personalized experience to returning users. However, it is possible to utilize cookies in a way that might eventually compromise the privacy of the users. Since all information stored in cookies is handled under the hood by the sites that originally created them, it can be hard to tell exactly what they are being used for. Still, cookies provide great power to Search Engine sites as far as recording any type of information about their users and even keeping those tracking activities secret through the use of careful encoding and encryption.

The first serious privacy breach by a search engine arose in August 2006. AOL accidentally published search queries from more than half a million users, made over a three-month period, on its website. By the time AOL realised its mistake, the information had been copied to a number of other Internet sites, which are still available today. In 2005, the U.S. Department of Justice subpoenaed Google, Yahoo, MSN, and AOL for tens of millions of users' search queries. Google successfully fought the request, and was able to limit its disclosure, but it is unknown how much data other companies may have turned over.

Like AOL, Google also stores users' search terms. This is done in the form of a server log. Each search creates a new entry in the log.

Structure of Google's Search Logs



```
203.14.90.34
23/02/2007 17.22.21
www.google.com/
search?q=linklaters
Firefox 2.0.0.3; Windows XP 3.1
589326gu8946e887
```

The server log information has greater potential to identify the individual than the information disclosed by AOL. It not only contains search terms but also traffic data, such as the IP address of the user and the user's Google cookie. This information could, in theory, be used to actually identify, rather than just guess the identity of, the user. Google collects additional information about users if they use other services such as Google Mail, Talk and Calendar. Of these, Google Mail has attracted the most controversy because of its use of "content extraction". This analyses incoming and outgoing e-mail in order to target the advertising to the user while they are using the Google Mail service. Google continues to use this technique on Google Mail accounts despite numerous privacy complaints from various organizations such as the Electronic Privacy Information Centre.

3.1 Double Click

The privacy concerns surrounding Google have been heightened recently by its plans to acquire DoubleClick, Inc., one of the leading providers of Internet advertising. In order to help its clients target the advertising it also allows users to be tracked as they visit different websites across the Internet. This is done by placing a DoubleClick cookie on the user's computer the first time they visit a site with a DoubleClick advert. This cookie is then read each time the user visits another website containing a DoubleClick advert, thus building up a picture of the user's surfing habits.

3.2 Google's My Search History

A new feature launched by Google allows users to see all of their past searches. The service, called My Search History, is similar to, but more comprehensive than, the feature Amazon.com, Ask, and America Online have offered for some time. It is intended to help people who use Google locate the information they sought during earlier searches so they can avoid repeating past queries. Once a user has set up the account, he or she will be able to see the search words previously entered as well as the sites visited previously that contain information on that search term. This may sound interesting and useful, but computer experts said there are risks to privacy the technology has now generated. By this, a user allowing Google to store search history on their computers and as long as Google holds up its end of the privacy policy, that information should remain safely on its servers. It is a universal truth that all search engines, including Yahoo, Google and MSN, retain search data of their users, which can easily give a clue about the person's identity and a glimpse into his mind and online activity.

4. Some Remedial Measures to Search Engine Privacy

Individuals concerned with privacy but wishing to use the valuable services of search engines can adopt a number of measures to safeguard their personal confidentiality. Before type a search terms into a search engine box or register for extra services at a search engine, it is necessary to be aware of the potential consequences. Searches can come back to haunt, especially if they are

Services Edition)

problematic and can be tied directly to users in some way. Here are some tips to help to enhance Web searching privacy ranging from high protection steps to simple steps we can start to take right away.

Watch what you search for: By avoid using terms that include full legal name attached to any information. For example, searching for users' full name and ID card number in a query is not optimal. If these types of search have conducted, then the users name and ID will appear together in the search string, and may be stored for a long time by the search engine.

Special note about passwords and user names: By avoid typing passwords or user names into search engines. If there is a security breach that allows users data to be released to others, these passwords and user names can potentially be used to identify the surfer, or even potentially cause some mischief. Therefore, it is a good idea to change passwords and user names, if the passwords or user names has been entered into a search engine.

Considering using an anonymizing tool or a proxy: The simplest way to disassociate from search terms is to use an anonymizing tool. There are free services available that allow using the Web without revealing users computer address, and there are also pay services. We may not realize it, but it is true that a computer disclose a lot of information as we traverse the Web. For example, login through IP ID at http://ipid.shat.net, we can verify someone computer or IP address and the kind of information that computer is disclosing.

TOR Onion Router and Privoxy: TOR (http://tor.eff.org) is a free tool that can be install in combination with a tool called Privoxy (http://www.privoxy.org/), which helps to mask yours computer's address, among other things. TOR and Privoxy are a good tool set and are well worth considering. These two tools should be used together.

Use Scandoo - Scandoo is a wonderful wrapper written around search engines that warns you of malicious websites in search results. Scandoo can help you search Google, Yahoo or MSN without disclosing your actual geographic location or IP address to the search engine. Scandoo interface remains invisible to the end user.

Download HideMyIp software - Your IP address is one big link between your search queries. If you are using a static IP, you can still hide it with HideMyIP address software. HideMyIP conceals your real IP address and shows a fake IP with a hostname to the sites that you visit. You can set Hide-My-IP to change your IP address every minute.

Download Customize Google for Firefox - If you Google using Firefox, this is a highly recommended extension that completely enhances your Googling experience. It can help remove Google Ads, anonymize your Google user ID, remove click tracking or filter Google search results.

Disallow Google to Store Cookies - The important thing is that it doesn't suffice blocking cookies from just google.com domain; you must also block cookies from google site in your country. For example, in India, one would block google.com and google.co.in. This is because Google redirects you to your local country page when you type in google.com in the browser address bar. To block cookies, open the Cookie blocking dialog in your browser, type the site URL and click disallow or block.

4.1 Some General Tips for Using Search Engines

These following tips are small steps that will not completely protect us from all search engine privacy issues, but they can potentially help to make incremental improvements.

- Do not accept search engine cookies. If you already have some on your computer, delete them. Cookies can be used to correlate a variety of information.
- Do not sign up for email at the same search engine where you regularly search. If you do so, then your email address can potentially be tied to your search terms. Whether or not a search engine does this is usually disclosed in the search engine's privacy policy.
- If you surf using a cable modem, or a static Internet connection, ask your service provider to give you a new IP address. Changing IP addresses every once in a while can be helpful for people who primarily surf the Web from one computer in one location over a long period of time.

5. Conclusion

It is challenging to achieve 100 percent privacy 100 percent of the time when using search engines due to the large amount of information the search engines collects. The World Wide Web, in its all-encompassing magnitude, has been responsible for new paradigms in several disciplines, such as communications, data management and individual privacies. Search engines are but one of the many elements to consider in this environment, where fundamental concepts of traditional law are extremely hard to apply. Perhaps the main conclusion that can be drawn from the conflict is that if every element involved acts on behalf of the best interest of privacy, avoiding extreme positions, eventually a reasonable status quo might be achieved. If individuals are extremely careful about revealing private information and make use of anonymizing tools wisely, if search engines never put business interests over privacy issues, and if governments do not overact on their attributes by disregarding individual freedoms in the name of national security, then a fair balancing of all the issues at stake could be achieved.

Many observers, reflecting on the importance of Search Engines today, are in fact considering the possibility of regarding them as a public service and legislating upon them accordingly, thus bringing a set of regulations that would provide a common frame on their operation including privacy issues such as those mentioned throughout this article. This type of international effort has already been attempted with varying degrees of success on other Internet security related areas. There are many alternative ways to find personal information online. It is clear that web users need to be extremely careful in keeping their personal information as secure and private as possible. However, many believe that it is possible that Search Engines may provide access to personal information for which there would be no other way to access, and are keeping a watchful eye over them. However, Search Engines provide a service and hence have detailed terms of service and privacy policies that people implicitly agree to when choosing to use it. Most Search Engines reserve the right to share aggregate information about the users and their search habits with third parties such as business partners and advertisers. This means that the personal information that may have been gathered about the search habits of individuals could be sold as part of the company to other corporations that might choose to make any type of use with it. Lastly, we can summarized that, it is up to the user to weigh the pros and cons of the prospects and then eventually choose to continue to use the services provided by the Search Engines.

References

1. Agarwal, Amit, "Google Privacy - What Data Do They Keep, How Long They Keep It", Available at < http://www.labnol.org/ > Site visited 20/09/2007.
2. Agarwal, Amit, "How to Stop Google from recording your Search habits?", Available at < http://www.labnol.org/ > Site visited 20/09/2007.
3. Arnold, S.E, "Who Can You Trust? Google, Yahoo or Overture", Available at <http://www.arnoldit.com/articles/yahoogooglemar2003.html> Site visited 20/03/2004.
4. Arnold, S.E., "Search: the New Application Platform", The Electronic Library, v.24 (2), 2006, pp.121-125.
5. Battelle, John. The Search: how Google rewrite the rules of business and transformed our culture. USA: Portfolio, 2005.
6. Bradley, Phil, "Librarians and Google: Tips of the Trade", Available at <http://blog.searchenginewatch.com/blog/060725-092829> Site visited 03/08/2007.
7. Edelman, Ben and Hannah, Rosenbaum, "The Safety of Internet Search Engines – Revisited", Available at <http://www. siteadvisor.com > Site visited 11/12/2006.
8. Ewalt, D.M, "Next-Generation Networks: the Evolution of Web Search", Available at <http://www.forbes.com> Site visited 18/05/2005.
9. "Google's Marissa Mayer on The Future of Search", Available at < http://

www.readwriteweb.com/archives/google_marissa_mayer_future_of_search.php> Site visited 28/06/2007.

10. http://www.comscore.com/
11. http://www.netratings.com/
12. Madden, A.D., Eaglestone, B., Ford, N.J. and Whittle, M, "Search engines: a first step to finding information: preliminary findings from a study of observed searches", Information Research, v. 12 (2), 2006, pp.294, Available at < http://InformationR.net/ir/12-2/paper294.html > Site visited 20/07/2007
13. Pandia Search Engine News, "The problem of search privacy – and some solutions", Available at < http://www.pandia.com > Site visited 21/07/2007.
14. Pandia Search Engine News, "Andrew Goodman on the future of Google and the search engine industry", Available at < http://www.pandia.com > Site visited 01/05/2007.
15. "Search engines, privacy and legal issues", Available at < http://www.cre8asiteforums.com/ > Site Visited 11/09/2007
16. Sullivan, Danny, "AOL Has "Safest" Results & Free Results Safer Than Paid", Available at http://searchengineland.com/061212-100923.php> Site visited 12/08/2007.

ABOUT AUTHORS

Prof Narendra Lahkar is the Head of the Department of Library and Information Science, Gauhati University. He has done his Mlib.Sc from Gauhati University, MSc in Information Studies from the University of Sheffield (UK) and Ph.D. from Gauhati University. He is the chairman of the Sub-Committee for drafting the Library Legislation of Assam. He is also the President, Assam Library Association, Zonal Secretary, Indian Association of Teachers of Library and Information Science (IATLIS). Presently he is supervising a numbers of research scholars for their Ph.D degree.

Mr Sanjib K Deka is presently working as a Sr. Library Information Assistant in the Central Library, IIT Guwahati. He holds MLISc, PGDCA and currently pursuing Ph.D. under the guidance of Prof. Narendra Lahkar, HOD, Library & Information Science Gauhati University.