World Applied Sciences Journal 20 (5): 753-759, 2012
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2012.20.05.261212

Search Engines in Website Security Leak

[1]Rizik M.H. Al-Sayyed, [2]Bader Al-Fawwaz, [3]Omar Al-Adwan,
[3]Hussam N. Fakhouri and [4]Mohannad S. Al-Khalaileh

[1]Department of Business Information Technology, The University of Jordan, Jordan
[2]Department of Computer Information System, Al-BAYT University, Jordan
[3]Department of Computer Information Technology, The University of Jordan, Jordan
[4]Department of Computer Science, The University of Jordan, Jordan

Abstract: Search engines should access certain website directories and specific information through queries; some of these queries can be used by hackers to access private sites information without the permission of their owners. Many websites have security leaks that allow search engines to penetrate into them and have access to critical information that are classified as private or secure. This paper aims at investigating the leaks in websites' secured information that the search engine can access through many queries, it also shows how these leaks and security gaps are prune to the danger of website security and shows the suggested solutions to avoid website vulnerability in order to overcome these issues.

Key words: Search engine · Robotics · Security leak in websites · Search queries · Website penetration

INTRODUCTION

One of the important services offered by the Internet is search engines; they allow users to get their needed information easily and quickly. Search engines navigate through the World Wide Web and get back to users so quickly with what they requested. Searching can only be performed through the sites that have been added to these search engines. People normally search through the web using queries; specifying these queries precisely and accurately help in limiting the results to the minimum range to simplify viewing and extracting information. Users are expected to learn some symbols and keywords to help them make a successful search query; for example in Google, people can use: +, -, " ", OR, intitle, allintitle, inurl … etc.

"82% of websites have had at least one security issue, with 63 percent still having issues of high, critical or urgent severity." (WhiteHat Security, 2008). "70% of the top 100 most popular Web sites either hosted malicious content or contained a masked redirect to lure unsuspecting victims from legitimate sites to malicious sites." (Websense, 2009)

Being on the top search engine results is very important for website owners. So, in order to promote a website among millions of others the website owner sometimes tends to index all pages and the content of these pages in the search engines in all means and methods available, some of these methods show the negative impact of these changes in promoting the website and lead to security leak.

The process of promoting the websites is called Search Engine Optimization (SEO); it is considered a basic technical issue that has a clear impact on the on-line marketing process; it is a clear fact such that the SEO helps in advertising electronic websites and increases the number of the website visitors. This requires little technical knowledge; such as the basic of the HTML Hypertext Markup Language. Sometimes, the SEO is called SEO Copyrighting because the process of preparing the websites for search engines basically deals with texts. It is possible to define the process of advertising the websites on SEO search engines as a preparation to web pages or even complete website to become closer to the search engines and this in turn, produces better results during the search process.

As being connected to the internet, the connected computer becomes at risk. This risk is due to many reasons including viruses and hackers. The solution is certainly not to turn your machine off. Websites vulnerability comprises two types: software and bad

Correspondig author: Rizik M. H. Al-Sayyed, Department of Business Information Technology, The University of Jordan, Jordan.

configuration. Most hackers penetrate easily into websites with known misconfiguration; the focus of this paper.

Owners of the websites should properly configure their websites in order to avoid illegal view and/or access; only viewable information should be viewed by search engine users and nothing else. Leaving important private information/files on the directory where search engines can navigate to is considered a leakage in website's security. When a search engine gets access to a website, it can view/index all those files in the search directory; this makes the content of these directories available to others without their owner's permission.

In this paper, we followed the empirical method to show the study points of leakage; we believe that by checking practically and experimentally these security issues, we can show the cases clearly. We also used the Google search engine in all the experiments due to its popularity and flexibility.

The rest of this paper is organized as follows:The next section introduces some literature review about the topic; it is followed by a section that covers the security issues with examples. We then present the potential solutions and the proposed model. The conclusion and future work is drawn in the last section.

LITERATURE REVIEW

Searching and hacking websites is a hot research topic that has a good deal of literature. A big number of websites that are indexed in search engines can be attacked by hackers who also can attach the network after collecting the needed information [1]. Due to the great number of hacking trials, it is difficult to track the IP address of hackers. People, in many cases help hackers when they don't care about leaving their computer connected to the Internet hours and days without watching what is being running on their computers and discover, lately, that they have been hacked.

The definitive source for information about Google Hacking is Long [2]. The field of ethical hacking, as opposed to malicious hacking, has grown significantly in the past 5 years due to the overwhelming demand for properly trained security professionals [3, 4].

Lancor and Workman [4] describe incorporating Google Hacking into a graduate course on web security. Billig, Danilchenko, Frank in there paper [5] described a good introduction to Google Hacking and a series of exercises used to teach students how to use Google Hacking to test their own sites and how to defend against it. And in there paper they looked at five of them: Error Messages, Open Directories, Documents and Files, Network Devices and Personal Information Gathering.

Johnny Long [6] descried the Penetration testing or ethical hacking is used for performing a security check on a website. Its aim is to try and find as many security loopholes as possible. Zai and Ayaz [7] in there project used Google search as a tool for penetration, the aim of their report is to examine such vulnerabilities by practically trying to penetrate a website and then suggest solutions to tackle them. They also implemented those solutions and tested them again to find how effective they are. They found that the implemented solutions were very effective and achieved success in preventing unauthorized access of information.

According to the Johnny Long's Google Hacking Database [8], there are roughly fourteen categories of Google hacks.

DISCUSSION

Leaving websites without or with poor configuration allow search engines to penetrate into them and have illegal access to critical and private data. We present in this section some of the security leaks that the search engines can penetrate.

Accessing Unprotected Database and Backup Files: Many website owners keep their backup and other files on the website in the format of xls, sql, mdb or as compressed files.zip or.rar on the website directory of the website files, these security gap occur due to the fact that these users are unknowledgeable that these files will be accessed by search engines such as Google. Website owners think that if the files are not linked to the websites pages, the other browsers (through search) will not see or access them while they are listed in the folders of the website. In fact the users are not aware that the search engine can penetrate into the directory and access these files and list them on the search engine hits returned during search. In addition, the information inside these files (content) can be viewed as well. In summary, anyone uses the search engine to search for these files can simply find and access them; these information make a big security leak. For example, in Google, the search query:

password filetype:xls

Returns about 53,900 results for accesses private information in Google search as shown in Figure 1. While exploring these files, users can find the ID and Password of users for many website that should be private and secure.

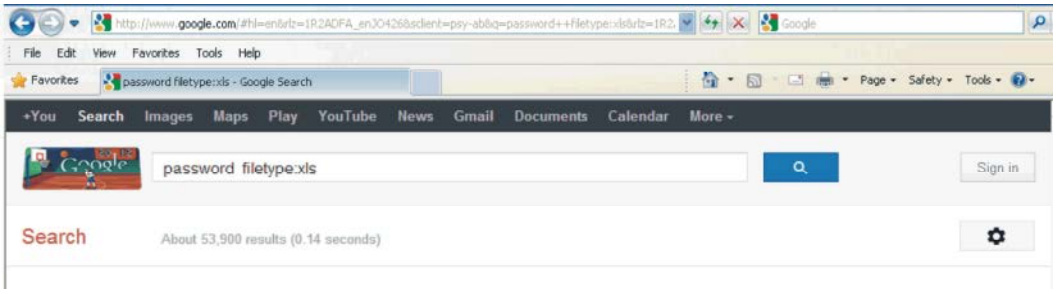

Fig. 1: Google search results for: password filetype:xls

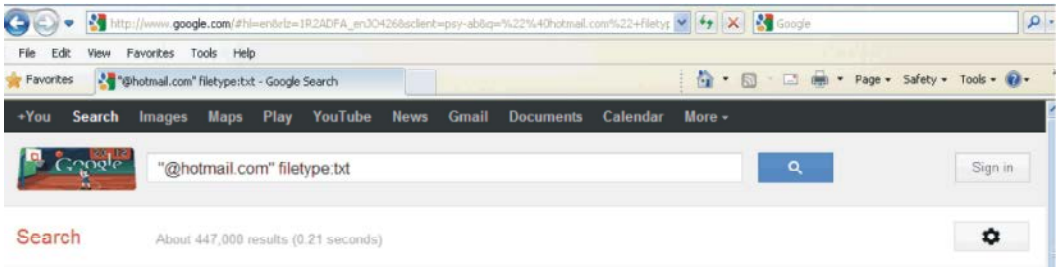

Fig. 2: Google search results for: "@hotmail.com" filetype:txt

Accessing Unprotected Mail List: Search engine queries can access information stored in the website directory and inside the text file or any other file stored insecurely on the server. Using some queries to search for emails, can lead to accessing private email lists that are private to companies and organizations. These email lists may be then sold, used by spammers or for any other illegal purpose.

For Example, in Google, the Search Query:

"@hotmail.com" filetype:txt

Returns about 447,000 results (as shown in Figure 2) for accesses to private mail lists and some of these lists may be very private and contain thousands of emails.

Accessing Secured Member Pages and Restricted Files: Some of website administrators who manage forum or social networks, index there secure pages in search engines in order to get more traffic to there websites, these pages are supposed to be accessed by member only, however, when indexing them by a search engine, they will bypass the security limitation by using ”cache” and it becomes insecurely accessible to public and this causes a security leak.

Another issue related to accessing member only area illegally and without the permission of the website owner happened on the search engine of the downloaded links that is related to some download manager software; when someone make a purchase and get the download link from the member only area, the download link will be listed in some mirror download search engine and this link become available to other users who can access the downloadable product directly without having to pay or getting to the member area and this security gap is due to the illegal indexing of these links in the search engine related to this download manager, which also can be listed in future by other search engines incase this website is also added to other search engines.

Accessing Unprotected Passwords and Credit Cared Information: Some search engine queries such as “allintext“ which locate the a text string within the body of the files found on the directory of the website or file pages, which also can find a string anywhere in the page except title, URL and internal/external links. Such queries can help in getting passwords or credit card information of the ecommerce websites that are unaware that by storing this critical information insecurely on the same directory of the website will make them available to search engines and also can be easily found by hackers; this causes a security major issue. An example in Google on this kind of search is shown in the following query:

Allintext: Credit Card Filetype:txt:

Accessing and Download Restricted Files: There are cases where PDF, DOC or similar files are accessible only to members and paid users or stored on the website directory as a storage place or further use, the search

engine robotics access these files on the website server and the user of the search engine, if aware of a small sentence in the document needed, can overcome this limitation by typing this sentence in the search query, if these files are unprotected properly, they can be download. For example, in Google, if the search query is included in double quotes and the filetype of the file is specified, the targeted results will be shown.

Generally you should avoid searching for Titles because they are used more often as a reference by irrelevant pages but if you know only the title of the document you can search for it by using the intitle and the filetype query. An example on this in Google:

Intitle:"Algorithm" Filetype:pdf:

If the owner of the website is not aware of protecting and securing their files and documents in the websites directory then the search engine can easily list these documents and the information inside them for public users.

Accessing Unprotected Internet Webcams: Installing web cameras and leaving it improperly configured allows web users to access it without the need for a username and a password. Indexing it by Google will even make the case worse by which viewing these pages allow intruders to spy on these cameras owners.

Examples on this in Google Are:

intitle:"Live View / – AXIS" | inurl:view/view.shtml^ inurl:/view.shtml

POTENTIAL SOLUTIONS AND THE PROPOSED MODEL

The searching process includes crawling, indexing and arranging results. All words in the search query are looked and arranged based on many factors including the relation between the word and the website and page to appear and the relation between the word and both the internal and external links. The search result lists the pages that are archived with the search engine (e.g. Google) not those in the web. Pages are archived either using the links that lead to these pages or by using the added links available in sitemaps. While searching, some words are ignored (such as a and, or, on) and the more key (basic) words in the page, the higher the rank of that page will be in the results; all results will be shown with higher rank on top.

The search engine creates an index for the information (pages) being read; it is this index that helps users to search for pages that matches their query. Searching is an extensive process that employs many computers (called Googlebot in Google) that are crawling the net all the time and the results of the search is produced by combining the results from all these computers. In order to promote a website, it must be part of the Googlebot (as named by Google).

The main purpose of any search engine is to help people find what they want. The process starts by specifying the needed information and then this need is converted into a query that will be passed to a search engine which in turn looks for matching web pages (or other formats) and collects and brings up the results. It is extremely important to know that the search engine will not be able to get the results if it is not stores, so the content, the architecture and the navigation through websites are highly important. Figure 3 illustrates how local search engines work.

As shown in Figure 3, the user fills the search form, the query parser then tokenizes the terms and looks for operators and filters passes the search to the query engine to find documents with matched criteria in the index file, the matched documents are then ranked according to many factors including locations and matched terms and then the formatted results will be displayed to the user. So, searching is a sophisticated job that preparing index ahead of time.

Sometimes, websites owners and due to some considerations, like to exclude some pages from being indexed. Control to how the website can searched is stored in a file called robots.txt. While searching, the search engine looks for this standard (robots.txt) in the root folder of the search directory. The robots. txt file contains instructions for the search engine on how to access the site; access to the site can be entire site level, individual directory level, specific type pages level or individual page level. The better the owner builds the robots.txt file the better control on searching will be.

The Following Is an Example of a Robots.txt File:

```
User-Agent: Googlebot
Disallow: /logs/
```

The Content of the File Is Interpreted as Follows: User-Agent: Googlebot indicates that what follows are instructions for the Googlebot only, Disallow: /logs/ is an instruction to the Googlebot not to access the log folder. More examples follow:

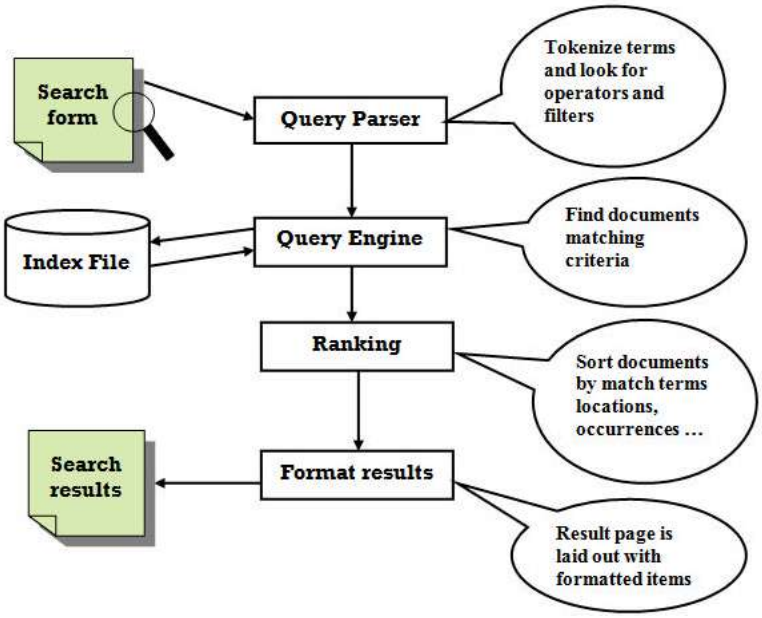

Fig. 3: Local Search Engine.

The Whole Site Can Be Excluded by Just Typing:

Disallow: /

A page can be excluded by typing (the name of the file is case sensitive):

Disallow: /my_private_page.html

An image can be excluded by typing:

User-agent: Googlebot-Image
Disallow: /images/family.jpg

A certain image type (e.g. TIF) can be excluded by typing:

User-agent: Googlebot
Disallow: /*.tif$

The robots.txt file can have other different rules. Different rules can be specific to different search engines.

To add more detailed rules on specific pages of the website, owners can employ the robots META tag. To control how an individual page is indexed, META tag is added to that HTML page. Both the robots. txt file and META tags increase the flexibility in implementing sophisticated access policies for pages in an easy way.

The Following Is an Example on Using Meta Tag:

```
<html>
<head>
<meta name="googlebot" content="noindex, nofollow"/>
...
</html>
```

In the above file, the line that uses META tag informs the web crawler that it should neither index the page nor follow any links on it.

Another Example:

<meta name="googlebot" content="nocache"/>

The above line instructs the search engine not to show a local cached copy of the page; this is useful for publishers to allow crawlers to index the entire contents and allows navigation to their pages.

The robots.txt file and the META tags are very beneficial in adding accessing policies to websites if they are used properly but both of them have limitations. Although the robots.txt file has the ability to add rules site-wide, is an optional file and needs to built properly and interpreted by search engines the way the owner of the website wants. The META tag is useful if the person

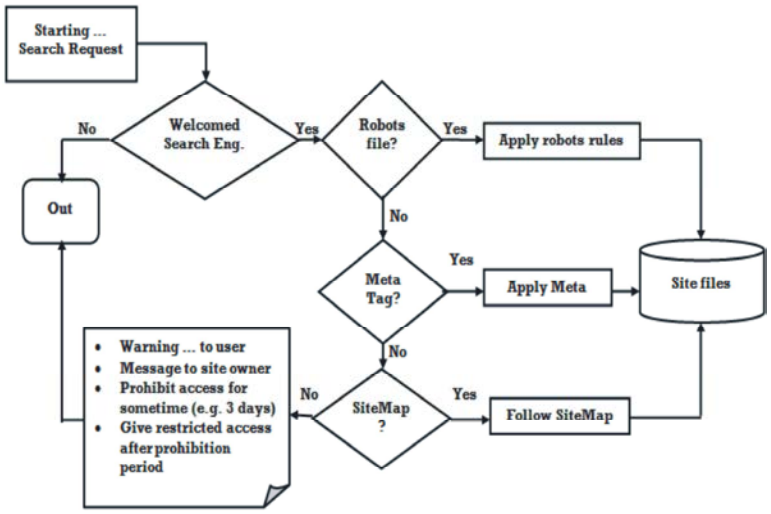

Fig. 4: Proposed Model.

managing the website has permission to edit the individual files and welling to edit the pages page by page.

Among Many Precautions That Should Be Fully Controlled by Websites, the Following Issues must Be Addressed:

- Administrators must differentiate between two types of pages: crawled by Googlebots and secret ones.
- Administrators can add in the public folder the file robots.txt; it is a text file used to exclude content from the spiders/bots of the search engines crawling.
- In order to stop search engines from caching a page, Meta tag can be added on each page.
- Robots Meta tags can be used to clear caches from files that can be used for unauthorized use.
- Data that is not needed shouldn't be stored on websites or on the accessible database; we should assume that there are always ways to reach the data.
- Prohibit access to servers; uploading files into servers could be dangerous as these files might have malicious payload.
- Enforce session authentication using proper methods such as passwords and keep track of access logs in a secure place while controlling illegal access that degrade the performance.
- Proper configuration and safe system and database structure information should be guaranteed. Bad or unwise configuration and/or revealing system's information put the server and the website at risk.
- Use available tools (such as SiteDigger) that can find week points in your website.
- In case of errors, only administrator should be informed, do not publish.
- Watch input carefully, validation should be strictly applied.
- Use additional security protocol such as: XML, JNOS, … etc
- Keep changing your IP address to make it difficult for people/sites/hackers who knew it.

To cope with the difficulties that owners of the websites might face due to bad configurations or any over sighted problem that any search engine might penetrate without the permission of the site's owner, we propose the simple model shown in Figure 4.

As shown in Figure 4, only eligible search engines can access site's files and unwelcomed ones, will be rejected and prohibited from accessing the website. Eligible search engines will first look for the robots.txt file, if it is there, its rules will be followed by the search engine assuming that all search engines obey the standard instructions. If the robots.txt is not available, however, the search engine continues looking for Meta tags implemented in individual files, if they are implemented, they will be obeyed. If neither robots.txt file nor Meta tags implementation are available in the website, the search engines keeps looking for the SiteMap as the minimal limitation needed for access restrictions, if it is available, only the pages navigable through it will be accessed, otherwise a warning to user will displayed informing him/her that access is not possible to the user *currently* followed by a message to be sent to the site's owner that the site has no rules to restrict searching, search through the site for limited period (e.g. 3 days) access to site and

after this period, users can access the website and owners understands the consequences risks of having some critical information available to others. We notice that having at least one of the three controllers (robots.txt, Meta tag, SiteMap) gives the user the ability to access the website and the more controllers implemented the more the site is secured if they are configured properly.

CONCLUSION AND FUTURE WORK

Search engines are important in bringing to users the information they need; if this information is available. The question that is always raised: are search engines positive tools always or could have drawbacks? During searching, search engines might reveal private or secret information that should be kept confidential. As people trust the search engines and gives them access to their websites, search engines must obey and respect owner's privacy. Speaking technically, there are a set of rules that kept in some kind of file controllers with instructions to search engines to obey. These instructions inform search engines for example to disallow access to some folders, files, images etc. As these controllers might not exist or badly configures by websites owners, we proposed a model to handle these shortages; in case of any. The model handles how both the search engine and the searched sites should interact to maintain websites privacy at a high level; of course, we assume high class search engines that respect privacy and the standards. A set of precautions is also presented to solve or mitigate the effect of free search that can reach all kinds of information wherever it is stored.

In our future work, we will implement the proposed model and try it with some common search engines.

REFERENCES

1. Hernández, J., C. Sierra, J.M. Ribagorda and A.B. Ramos, 2001. Search Engines as a Security Threat. Computer, 34(10): 25-30.
2. Long, J., 2008. Google Hacking for Penetration Testers, Syngress Press, pp: 2.
3. Palmer, C., 2001. Ethical Hacking. IBM Systems Journal, 40: 3.
4. Lancor, L. and R. Workman, 2007. Using Google Hacking to Enhance Defense Strategies. SIGCSE Bull, 39(1): 491-495.
5. Billig, Danilchenko, Frank, Evaluation of Google Hacking. SIGCSE Bull. Conference'08, September, 2008, September 26-27, 2008, Kennesaw, GA,USA.
6. Johnny Long, C.S., 2005. Google hacking for penetration testers, http:// www.blackhat.com/ presentations/ bh-usa- 05/bh-us-05-long.pdf, 19th April 2009.
7. Azam Zai and Muhammad Ayaz, 2009. Website Penetration via Google Search, 28th, April 2009.
9. Google Hacking Database Web Site, http:// johnny.ihackstuff.com/ ghdb.php.*password filetype:xls*.