SearchOL: An Information Gathering Tool

Farhan Ahmed[1], Pallavi Khatri[1](✉), Geetanjali Surange[1], and Animesh Agrawal[2]

[1] ITM University, Gwalior, India
{Pallavi.khatri.cse,Geetanjali.surange}@itmuniversity.ac.in
[2] NFSU, Gandhinagar, India
animesh.agrawal_dc@nfsu.ac.in

Abstract. Most organizations and their network administrators are familiar with penetration testing and possible attacks that can be done on a system through any software or hardware vulnerabilities of the system. System administrators, however neglect the quantity of system and user information that can be extracted anonymously from the content that is publicly available on the internet. This publicly available information is critical and of great use to penetration testers who wish to exploit the system. This work proposes a tool called 'SearchOL' developed in Python for gathering user related data from social sites using multiple search engines. Tool collects data passively and from the results proves to be a comprehensive data aggregator from multiple social platforms. The tool can be used for Information gathering which is the first phase of ethical hacking. The novelty of this tool is that it gives most important, most relevant and concise results from various search engines. It will help in reducing the efforts of the pen testers to gather information from public domains.

Keywords: Ethical Hacking · Reconnaissance · Penetration Testing · Vulnerabilities

1 Introduction

Now a days, everyone using social media and update their daily activities on social media. Autofill forms that are filled while creating any account on any social site store lots of personal information of an individual. We unknowingly provide lot of sensitive information to know people about out likes, dislikes, location, friends that can be exploited if gathered by attackers. Foot printing and reconnaissance process is used by hackers for gathering sensitive information from social sites and online presence. It is also used by ethical hackers to check that sensitive information is discoverable from the social sites and suggest techniques to hide it. Many tools are available online for gathering information having different functionality and use cases. Reconnaissance can be done in two ways:

1. Passive
2. Active

A. Abraham et al. (Eds.): ISDA 2022, LNNS 716, pp. 343–349, 2023.
https://doi.org/10.1007/978-3-031-35501-1_34

Passive reconnaissance means gathering of information without the direct interaction or connection with target. It can be done from social sites and internet searching. Social engineering is also considered as passive information gathering method.

In Active reconnaissance we directly interact with target network or system for gathering information. This has high risk of detection than passive reconnaissance. It involves discovering of hosts, IP addresses, services, router on the network.

Proposed tool 'ShearchOL' decrease the efforts done by ethical hackers for doing passive reconnaissance about target by simply entering the keyword (person name or organization name) and the tool will search the most relative informative websites links from Google, Ask, Yahoo, Bing search engines and save the links in a text file for further analysis. This decreases the overhead that ethical hacker faces by going one by one on different search engines and search information about the target. This is automatically done by proposed tool 'SearchOL'.

This article is organised in to 6 sections where the concept in introduced in Sect. 1, survey of literature has been summarised in Sect. 2. Section 3 and 4 brief about the Proposed work and experimental setup to conduct the experiment. Results and discussed in Sect. 5 and the complete work is concluded in Sect. 6.

2 Literature Survey

The process of ethical hacking starts with the gathering of more and more information about the target, we can gather simple and sensitive both information from social media and internet sites, the gathering of information about the target is known as Footprinting. There are many tools available to do footprinting. With the help of Footprinting we can gather information about network such as Network ID, domain name, IP address, protocols, news articles, web server links etc. If hacker get some very sensitive information, he or she can use this information for its malicious activities [1]. Authors in [2] proposes a cyber-reconnaissance tool named SearchSimplified built using Java. This tool gather data related to the organization entered. This tool gather data using Google's cache system, advanced query operators such as intitle:, site: filetype:. This tool gather data with the help of Google. Work proposed in [3] provides survey and taxonomy of adversarial Reconnaissance Technique, this paper tells us about cyber kill chain, Open-Source Intelligence, Sniffing.

Cyber Deception and case studies of cybercrimes, categories of Target information for reconnaissance, external an Internal Reconnaissance, Taxonomy of reconnaissance techniques, Defensive Measures against Reconnaissance Techniques, etc. In paper [4] author shows various web-based platforms for collecting and tracking IP information. Author performed an experiment in specialized university computer lab, Connect all hosts machine in the lab in Local Area Network (LAN). The results provide host name, Autonomous system, Internet Service Provider, country, continent etc.

A Comparative Study on Web Scraping is done in [5]. In these various practices of web scraping is shown. The author shows the multiple web scraping techniques by we can easily scrap websites, and compare various web scraping software. This paper gives the knowledge of various web scraping tools and techniques. We can easily and efficiently gather data from publicly available sources using OSINT (Open-Source Intelligence) tools of OSINT used in investigation phase for collecting information about target. The use of OSINT to gather information is shown in [6]. The proposed work uses the API keys of social media platforms and python libraries to check usernames exists or not, if exists gather data, store the results in database and display results in UI. The outcome of this study offers a review on web scraping techniques and software which can be used to extract data from web sites.

3 Proposed Work

From the extensive literature survey, following conclusions are derived:

- Existing system cannot Search using multiple search engines.
- They only search the usernames in different social sites.
- They require multiple dependencies to be installed before data searching.
- They search data from Organizations only and not from social accounts.

To extract precise and more information from the web this work proposes a Python based web scraping tool called SearchOL that will work on Google, Bing, Yahoo, Ask and will retrieve most relatable URLs from various websites.

The tool will also store the retrieved information in a text file that can be further used by an attacker to exploit the system or a user. The proposed tool used Advanced Search technique of Google search engine called Google Dorking [7] to discovers the data.

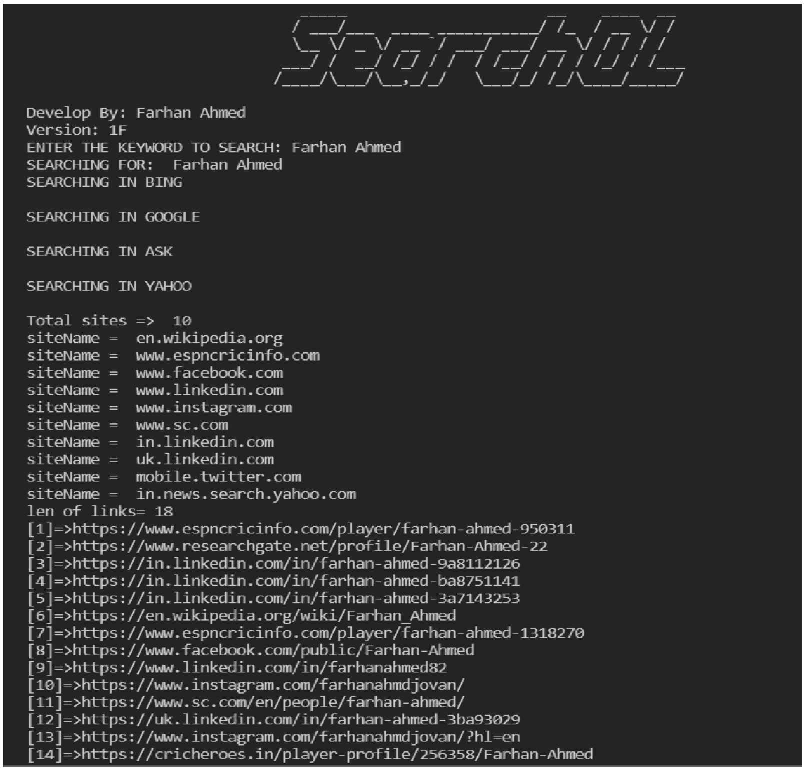
The working methodology of SearchOL is as follows:
It uses Python Requests Module [10], that allows to send HTTP requests using Python, we use the requests.get (url) [10] method to send a GET request to the specified url and returns a Response Object that contains the server's response to the HTTP request. It uses Python library Beautiful Soup for parsing structured data. It allows to interact with HTML in a similar way to how you interact with a web page using developer tools.

It uses OS module [12] of python that provides functions for interacting with the operating system. This module is used to save the information that is gathered gather from search using SearchOL.
The work flow of SearchOL is described below:

- Take input
- Create url for the input keyword
- Make request for the url using Requests Module [10], one by one on Google, Bing, Yahoo, Ask Search engines [13].
- Find all links in the search result using Beautiful Soup [11] module.
- Filter the most related and useful links
- Append sites in a list named ‘sitelist’
- Append the links in a list named ‘links’
- Iterate through the ‘sitelist’ and ‘links’ to print the information
- Now, save the information gather in a text file using OS module [12].

4 Experimental Setup

The tool developed in Python version 3.9.7 [8] is tested on system with configuration: Processor: Intel(R) Core (TM) i5-10210U CPU @ 1.60 GHz 2.11 GHz, System type: 64-bit operating system, x64-based processor, RAM: 8.00 GB with search keyword as Farhan Ahmed and scraping is done from Google, Bing, Yahoo, Ask as can be seen in Fig. 1.



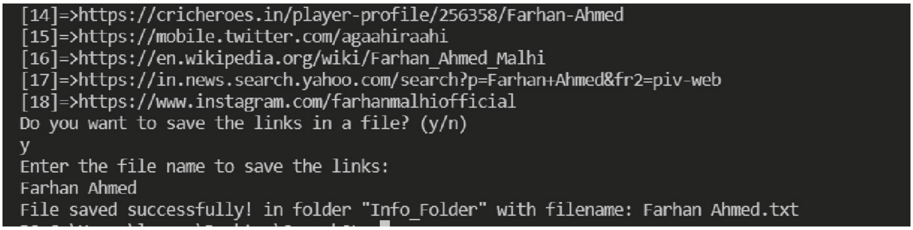
```
Develop By: Farhan Ahmed
Version: 1F
ENTER THE KEYWORD TO SEARCH: Farhan Ahmed
SEARCHING FOR:  Farhan Ahmed
SEARCHING IN BING

SEARCHING IN GOOGLE

SEARCHING IN ASK

SEARCHING IN YAHOO

Total sites =>  10
siteName =  en.wikipedia.org
siteName =  www.espncricinfo.com
siteName =  www.facebook.com
siteName =  www.linkedin.com
siteName =  www.instagram.com
siteName =  www.sc.com
siteName =  in.linkedin.com
siteName =  uk.linkedin.com
siteName =  mobile.twitter.com
siteName =  in.news.search.yahoo.com
len of links= 18
[1]=>https://www.espncricinfo.com/player/farhan-ahmed-950311
[2]=>https://www.researchgate.net/profile/Farhan-Ahmed-22
[3]=>https://in.linkedin.com/in/farhan-ahmed-9a8112126
[4]=>https://in.linkedin.com/in/farhan-ahmed-ba8751141
[5]=>https://in.linkedin.com/in/farhan-ahmed-3a7143253
[6]=>https://en.wikipedia.org/wiki/Farhan_Ahmed
[7]=>https://www.espncricinfo.com/player/farhan-ahmed-1318270
[8]=>https://www.facebook.com/public/Farhan-Ahmed
[9]=>https://www.linkedin.com/in/farhanahmed82
[10]=>https://www.instagram.com/farhanahmdjovan/
[11]=>https://www.sc.com/en/people/farhan-ahmed/
[12]=>https://uk.linkedin.com/in/farhan-ahmed-3ba93029
[13]=>https://www.instagram.com/farhanahmdjovan/?hl=en
[14]=>https://cricheroes.in/player-profile/256358/Farhan-Ahmed
```

Fig. 1. Output results of ‘SearchOL’

All information gathered from running the tool can be saved in a text file as shown in Fig. 2. The gathered information is vital as any penetration tester can use this information to exploit the target system.

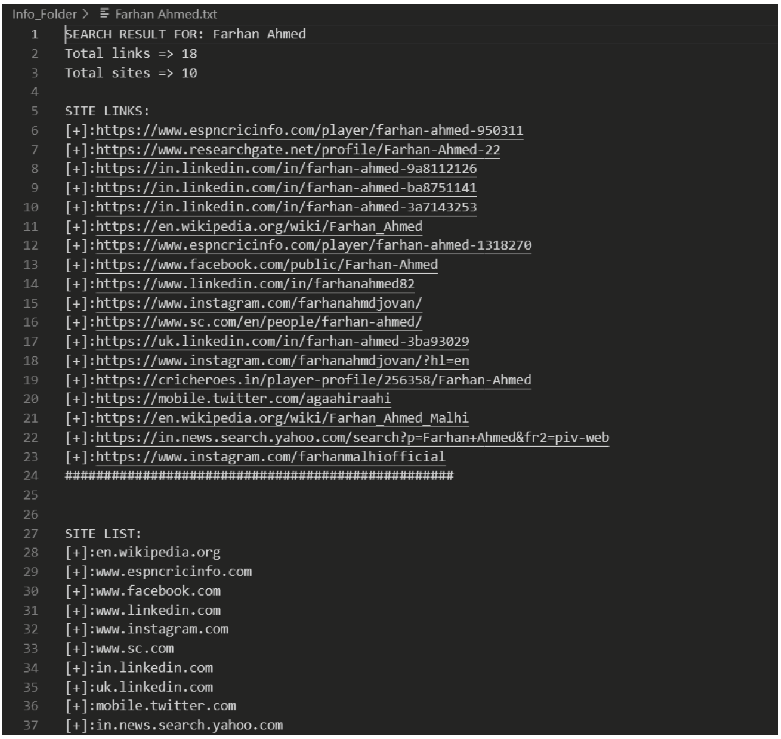
5 Results

A sample output from the tool is displayed in Fig. 3. As the user enters the name of an individual and proceeds to search on the social websites. The tool lists all the findings and allows to store complete data in a text file.



```
[14]=>https://cricheroes.in/player-profile/256358/Farhan-Ahmed
[15]=>https://mobile.twitter.com/agaahiraahi
[16]=>https://en.wikipedia.org/wiki/Farhan_Ahmed_Malhi
[17]=>https://in.news.search.yahoo.com/search?p=Farhan+Ahmed&fr2=piv-web
[18]=>https://www.instagram.com/farhanmalhiofficial
Do you want to save the links in a file? (y/n)
y
Enter the file name to save the links:
Farhan Ahmed
File saved successfully! in folder "Info_Folder" with filename: Farhan Ahmed.txt
```

Fig. 2. Saving all links in a text file.



```
SEARCH RESULT FOR: Farhan Ahmed
Total links => 18
Total sites => 10

SITE LINKS:
[+]:https://www.espncricinfo.com/player/farhan-ahmed-950311
[+]:https://www.researchgate.net/profile/Farhan-Ahmed-22
[+]:https://in.linkedin.com/in/farhan-ahmed-9a8112126
[+]:https://in.linkedin.com/in/farhan-ahmed-ba8751141
[+]:https://in.linkedin.com/in/farhan-ahmed-3a7143253
[+]:https://en.wikipedia.org/wiki/Farhan_Ahmed
[+]:https://www.espncricinfo.com/player/farhan-ahmed-1318270
[+]:https://www.facebook.com/public/Farhan-Ahmed
[+]:https://www.linkedin.com/in/farhanahmed82
[+]:https://www.instagram.com/farhanahmdjovan/
[+]:https://www.sc.com/en/people/farhan-ahmed/
[+]:https://uk.linkedin.com/in/farhan-ahmed-3ba93029
[+]:https://www.instagram.com/farhanahmdjovan/?hl=en
[+]:https://cricheroes.in/player-profile/256358/Farhan-Ahmed
[+]:https://mobile.twitter.com/agaahiraahi
[+]:https://en.wikipedia.org/wiki/Farhan_Ahmed_Malhi
[+]:https://in.news.search.yahoo.com/search?p=Farhan+Ahmed&fr2=piv-web
[+]:https://www.instagram.com/farhanmalhiofficial
##################################################



SITE LIST:
[+]:en.wikipedia.org
[+]:www.espncricinfo.com
[+]:www.facebook.com
[+]:www.linkedin.com
[+]:www.instagram.com
[+]:www.sc.com
[+]:in.linkedin.com
[+]:uk.linkedin.com
[+]:mobile.twitter.com
[+]:in.news.search.yahoo.com
```

Fig. 3. Saved text file

As summarised in Table 1 SearchOL tool has been compared with existing tools doing the same kind of work and results prove that amount of information that can gathered using SearchOL is more compared to others. This makes SearchOL tool more usable in case of information gathering about a person.

Table 1. Comparison of 'SearchOL' and other tools

Tools/Techniques	Source of data	Type of data	Restrictions
Sherlock [9]	Social media sites	Usernames	Only Usernames
Google Dorking [7]	Google	Files and sites	Only work with google
SearchSimplified [2]	Google	Only Organizational data	Only get organizational data
SearchOL (proposed)	Google, Bing, Yahoo, Ask	All links related to person or organization	None, it works with all search engines, can get organizational data as well as persons

6 Conclusions

As we see above that so much information is available on internet and this information if use for wrong purposes it will costs a lot. Many fraud calls, schemes, OTP frauds, Bank frauds done by just using your small-small information available online. Many hackers make fake accounts of victim user to deface him/her. The information we think useless, but it can have great impact on our life if goes in wrong hands. This paper provides the python-based tool 'ShearchOL' to gather important links related to the input keyword from Google, Ask, Bing and Yahoo search engines and save them easily in a text file for further analysis. The tool we present can be used by penetration testers to look for the sensitive information released on internet. So that they can take appropriate measures to protect the sensitive information. More features can be added to the tool for Data gathering. More search engines [13] can be added from where data can be scraped, secure web browsing [14] can be taken up as future work.

References

1. Shreya, S., Kumar, N.S., Rao, K., Rao, B.: Footprinting: techniques, tools and countermeasures for footprinting. J. Crit. Rev. **7**, 2019–2025 (2020). https://doi.org/10.31838/jcr.07.11.311
2. Roy, A., Mejia, L., Helling, P., Olmsted, A.: Automation of cyber-reconnaissance: a Java-based opensource tool for information gathering. In: 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 424–426 (2017). https://doi.org/10.23919/ICITST.2017.8356437
3. Roy, S., et al.: Survey and taxonomy of adversarial reconnaissance techniques. ACM Comput. Surv. (2022)
4. Boyanov, P.Kr.: Implementation of the web based platforms for collecting and footprinting IP information of hosts in the computer network and systems. Space Research and Technology Institute-BAS, Bulgaria Konstantin Preslavsky University-Faculty of Technical Sciences Association Scientific and Applied Research, vol. 16, p. 42 (2019)
5. Sirisuriya, S.C.M.de.S.: A Comparative Study on Web Scraping (2015)

6. Sambhe, N., Varma, P., Adlakhiya, A., Mahakalkar, A., Nakade, N., Lakhe, R.: Using OSINT to gather information about a user from multiple social networks. Inf. Technol. Ind. **9**(2), 207–211 (2021)
7. Parmar, M.: Google Dorks -Advance Searching Technique (2019). https://doi.org/10.13140/RG.2.2.24202.62404
8. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)
9. Sherlock-project. https://github.com/sherlock-project/sherlock
10. Chandra, R.V., Varanasi, B.S.: Python Requests Essentials. Packt Publishing Ltd. (2015)
11. Richardson, L.: Beautiful soup documentation. Dosegljivo (2007). https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Accessed 7 July 2018
12. Pilgrim, M.: Exceptions and file handling. In: Dive Into Python, pp. 97–120. Apress, Berkeley, CA (2004)
13. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information Retrieval in Practice, vol. 520, pp. 131–141. Addison-Wesley, Reading (2010)
14. Tang, S.: Towards secure web browsing. University of Illinois at Urbana-Champaign (2011)