A Survey on Cyberspace Search Engines

Ruiguang Li[1,2(✉)], Meng Shen[1], Hao Yu[1], Chao Li[2], Pengyu Duan[1], and Lihuang Zhu[1]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
lrg@cert.org.cn

[2] National Computer Network Emergency Response Technical Team/Coordination, Center of China, Beijing, China

Abstract. This paper introduces the concept of cyberspace search engine, and makes a deep survey on 5 well-known search engines, say Shodan, Censys, BinaryEdge, ZoomEye and Fofa, by querying official websites, analyzing APIs, and making academic research. We discuss the following items in details: Supporting internet protocols, Total amounts of detected devices, Device information, Scanning frequency, System architecture, The third party databases, Probes distribution, etc. We give a comprehensive comparison of the detecting abilities and working principles of the cyberspace search engines.

Keyword: Cyberspace search engines

Cyberspace search engines, such as Shodan, Censys, BinaryEdge, ZoomEye and Fofa, are new Internet applications in recent years. They search various types of online devices in cyberspace, such as webcams, routers, intelligent refrigerators, industrial control devices, etc. They are becoming powerful tools to detect network resources. At present, mastering the network resources is valuable for cyberspace governance and network security protection. Therefore, global security companies and scientific research institutions pay great attention on the development and utilization of cyberspace search engines. This paper will carry out a comprehensive investigation and analysis on the detection capabilities and working principles of 5 well-known search engines.

1 Introduction

Network resources exploration is to send probe packets to the remote network devices, and to receive and analyze the response data, so as to get the information of remote devices, such as opening ports and services, operating systems, vulnerability distribution, device types, organizations, the geographical position, and so on. The detecting protocols are mainly on the transport layer and the application layer in the TCP/IP stacks. The detection methods of transport layer include SYN scan, TCP connection scan, UDP scan, FIN scan, ICMP scan, etc. Application layer detection mainly uses the special fields of internet protocols, special files, hash values, certificates, and so on.

The working principles of cyberspace search engines are very different from the Web search engines such as Google, Baidu. Web search engines collect, store and analyze

W. Lu et al. (Eds.): CNCERT 2020, CCIS 1299, pp. 206–214, 2020.
https://doi.org/10.1007/978-981-33-4922-3_15

Web page for information querying, while the cyberspace search engines adopt the network resource detecting technology. By sending the detection packet to the remote devices, it can obtain the important information of the target, and conduct comprehensive analysis and display. Global security companies and research institutions have developed a number of search engines, in which the following are most well-known: Shodan (www.shodan.io) Censys (Censys.io) from the US, BinaryEdge (www.binaryedge.io) from Europe, and ZoomEye (www.zoomeye.org) Fofa (www.fofa.so) from China. Some of these engines are commercially available, while others offer none-profit services.

We are very interested in the detection abilities and the working principles of these search engines, so we made a comprehensive investigation on Shodan, Censys, BinaryEdge, ZoomEye, Fofa, by querying official websites, analyzing APIs, and making academic research. The main contents include: Supporting internet protocols, Total amounts of detected devices, Device information, Scanning frequency, System architecture, The third party databases, Probes distribution, etc.

2 Supporting Internet Protocols

Mastering various types of Internet protocol formats is the basis for the exploration of cyberspace search engines. Different devices in the internet have different protocols. In order to facilitate the comparative study, we first carry out a classification of various network devices.

We got all types of devices from the search engine's official websites, and classify all devices into 11 categories: Network Equipments, Terminal, Server, Office Equipment, Industrial Control Equipment, Smart Home, Power Supply Equipment, Web Camera, Remote Management Equipment, Blockchain, Database, shown as Fig. 1.

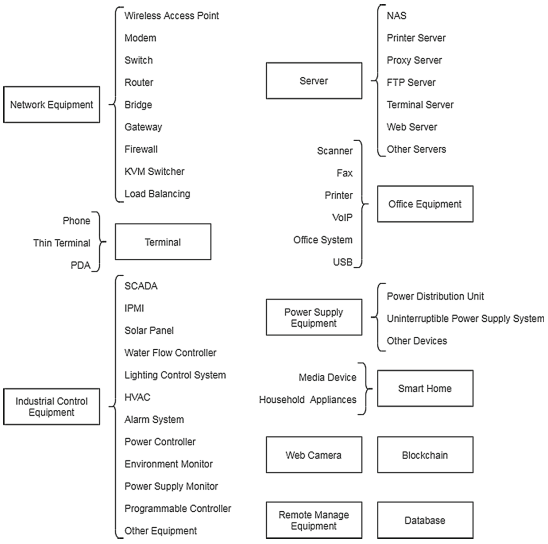

Fig. 1. Device categories

On this basis, we obtained the lists of all engines' supporting protocols from the official websites, user manuals, the APIs, and some technical forums. We classify them into 11 categories according to Fig. 1, shown as Table 1, where "-" means there is no such agreement.

Table 1. Supporting internet protocols

	Shodan	Censys	ZoomEye	Fofa	BinaryEdge
Network equipment	10	1	54	7	8
Terminal	19	1	227	6	13
Server	67	10	154	20	63
Office Equipment	12	5	31	6	11
Industrial Control Equipment	26	5	16	23	17
Smart Home	9	–	3	7	9
Power Supply Equipment	4	1	3	2	4
Web Camera	3	–	8	–	3
Remote Management Equipment	13	5	31	8	11
Blockchain	5	–	4	21	4
Database	17	6	19	16	15
Total	185	34	550	116	158

Shodan's API interface contains supporting protocols that can be directly queried [1]. Censys's protocols information comes from the official forum [2]. ZoomEye's protocols information comes from the NMAP-Services file in the user's manual [3]. Fofa's protocols information comes from the technical forum [4]. BinaryEdge's protocols information comes from the API documentation [5]. As you can see in the table, Shodan and ZoomEye have mastered more types of network protocols, covered all protocol categories, and presumably have better device detecting capabilities. Due to the different statistical caliber of network protocols, there may be some deviation in the comparison results.

3 Total Amounts of Detected Devices

Based on the analysis in Sect. 2, we investigate the total numbers of detected devices of different search engines. Typically, the official websites will claim the total numbers of detected devices, but sometimes we need to do more auxiliary analyzing.

The total amount of Shodan comes from the official website query tool CLi.shodan.io [6]. All the data records after January 1, 2009 can be inquired by the command line tool, so we can calculate the total number of detected devices.

The official website of Censys provides data statistics function [7]. We divide the IPv4 address space into 256 parts, and retrieve each address block with Censys, and calculate the manufacturer's brands of specific types in the returned results, and then obtain the total number as a summary. The total amount of ZoomEye, Fofa and BinaryEdge are from the official website [5, 8, 9].

Table 2. Comparison of the total amount of detectable devices

	Shodan	Censys	ZoomEye	Fofa	BinaryEdge
Total amounts	436489751	111368143	1190860679	270363	89871839

The total numbers of detected devices for each engine are shown in Table 2. As you can see from the table, ZoomEye (nearly 1.2 billion) and Shodan (over 0.4 billion) have the strongest detecting capabilities.

It should be noted that, because of the lack of industry standards in the field of network devices classification, there are statistical caliber of the comparison results.

4 Device Information

Cyberspace search engines need to present the detected device information in a comprehensive way for users to use. One device stands for a file or a record. By analyzing the files or the records, we can get the device information architecture. Typically, the device information architecture includes such important information as domain names, opening ports, services, geographic locations, countries, device types, affiliation, and so on.

We collect, analyze and draw the device information architecture of the above search engines, and make a comparison. We can classify all the device information into: Equipment information, location information, port information, loopholes, probe point information, tag information, network equipment information, WEB information, file transfer, email protocol information, remote access to information, database information, industrial control protocol information, message queues, clustering information. This will be of great value to developers and users of the cyberspace search engines.

Taking Censys as an example, by analyzing the official documents of Censys [10], we get the tree diagram of Censys' device information architecture, as shown in Fig. 2. All these information will be reflected on Censys' web pages. In the below figure, the vulnerability information and probe point information are represented as dotted lines because Censys does not provide such information.

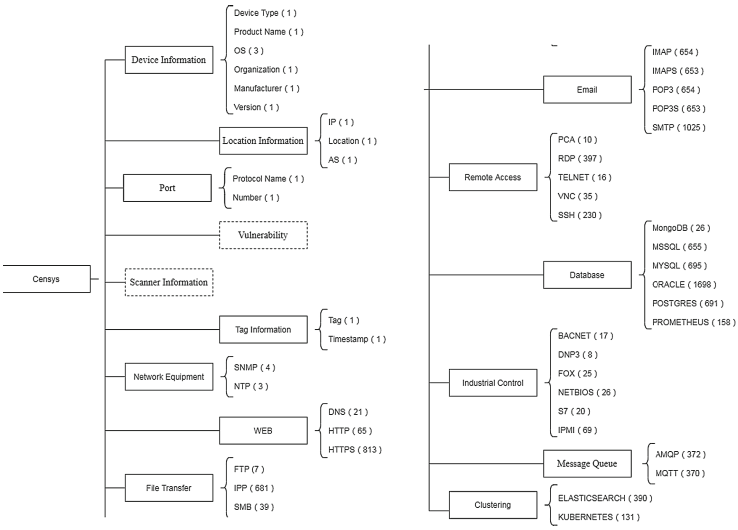

Fig. 2. Device information architecture for censys

5 Scanning Frequency

The cyberspace search engines constantly scan and probe the whole network, discover the new connected devices, and periodically update the detected devices. As a complete scan of the whole network consumes lots of computing and storage resources, so search engines usually set a scanning frequency. Scanning frequency is an important index for the detecting ability. The higher the frequency, the stronger the search engines' performance.

We measured the scanning frequencies of Shodan, Censys, ZoomEye and Fofa. More than 130 IP addresses (opening HTTP, HTTPS, TELNET, FTP and SSH services) were randomly selected. By checking the update status of these IP addresses every day, we can get the scanning intervals of each engines, as shown in Table 3 below.

Table 3. Comparison of scanning frequencies

Protocol (port)	Shodan	Censys	ZoomEye	Fofa
HTTP (80/TCP)	10 days	2 days	389 days	39 days
TELNET (23/TCP)	24 days	2 days	-	-
HTTPS (443/TCP)	9 days	1 day	26 days	102 days
FTP (21/TCP)	13 days	2 days	173 days	74 days
SSH (22/TCP)	10 days	3 days	24 days	60 days

In the above table, "-" means it hasn't been scanned for a long time. As can be seen from the table, that the scanning frequencies of Shodan and Censys are significantly

higher than that of ZoomEye and Fofa. We can include that Shodan and Censys have more powerful performance.

6 System Architecture

We are very interested in the system architectures of the cyberspace search engines, so we conducted an extensive academic research. Typically, the architecture of search engine can be divided into three modules: information acquisition module, data storage module and information retrieval module. The information acquisition module is responsible for collecting the information of various devices in the cyberspace. The data storage module is responsible for storing the massive device information collected, and the information retrieval module is responsible for providing statistical and querying services.

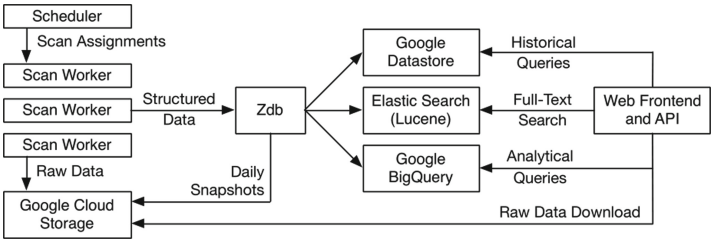

Fig. 3. Censys system architecture1

Figure 3 shows the system architecture of Censys [11]. In the above figure, the Scan Worker is responsible for information acquisition. The Scheduler allocates scanning tasks to multiple scanning modules. The scanning module will save the detection results to Zdb database, and all the information will be stored in Google Cloud. In the information retrieval module, Censys provides elastic Search for full-text retrieval. Google Datastore offers history retrieval and Google BigQuery offers statistics retrieval.

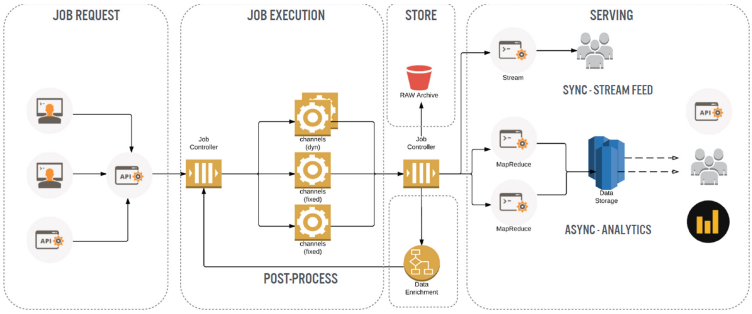

Fig. 4. BinaryEdge system architecture2

BinaryEdge system architecture is shown in Fig. 4 [12], which is divided into four parts: task submission, task execution, storage and service. Task submission uses HTTP,

command line, third party and other forms of API for data acquisition. In the task execution stage, the task is sent to multiple channels, including port scanning, Screen shotter, OCR and other technologies. In the storage stage, the collected information will be divided into original data and processed data, and stored in the database. During the service stage, the processed data will be sent to users through a real-time information flow, or deeply analyzed by MapReduce, Kibana, or InfluxDB.

7 Third Party Databases

Many cyberspace search engines work with third-party databases, such as IP databases, domain name databases, and geographic location databases. We investigated the third-party databases associated with commercial search engines, as shown in Table 4 below:

Table 4. Search engines associate third-party databases

	Shodan	Censys	ZoomEye	Fofa	BinaryEdge
IP database	Randomly generated	Randomly generated	-	-	-
Domain database	-	Alexa	-	-	Passive DNS
Address database	-	GeoIP	IPIP	GeoIP	GeoIP

In the table, the IP addresses of Shodan and Censys are randomly generated and do not rely on the third-party IP database. We haven't found the information of ZoomEye, Fofa and BinaryEdge. As for the domain name database, Censys used the domain datas provided by Alexa Top 1 Million Websites, while BinaryEdge used the passive DNS resolution service. We haven't found the information of Shodan, ZoomEye and Fofa. As for geographic location databases, Censys, Fofa and BinaryEdge all use the database of GeoIP, while ZoomEye uses the database of IPIP.net.

8 Probes Distribution

Cyberspace search engines often need to deploy many probes because there are many security devices (such as firewalls) in cyberspace, making it difficult to detect the network edges. Only by deploying widely distributed probes, can we minimize the impact of security devices and find more edge nodes as possible.

We conducted an extensive research, focusing on the open-source tools and third-party organizations. GreyNoise and BinaryEdge have done well.

GreyNoise is a tool for collecting and analyzing scanning traffics [13]. It found the probes of 96 search engines, including Shodan, Censys, BinaryEdge and ZoomEye, as shown in Table 5 below.

Table 5. Probes distribution marked by GreyNoise

	Shodan	Censys	BinaryEdge	ZoomEye
United States	31	398	368	-
Canada	-	-	37	-
Britain	1	-	236	-
Netherlands	10	-	86	-
Iceland	2	-	-	-
Romania	1	-	-	-
Greece	-	-	1	-
Germany	-	-	239	-
India	-	-	29	-
Singapore	-	-	27	-
Japan	-	-	-	16

BinaryEdge recorded the contents of received packets(including IP, ports and payloads) which it received by deploying honeypots all around the world. Because the honeypots do not actively interact with other devices, the data received in the honeypots are most likely send by the proves. Table 6 shows the global probe distribution of Shodan, Censys and BinaryEdge recorded by BinaryEdge during a period of 2000 days.

Table 6. Proves distribution marked by BinaryEdge

	Shodan	Censys	BinaryEdge
The United States	17	321	146
Canada	-	-	24
The British	1	-	90
In the Netherlands,	11	-	36
Iceland	2	-	-
Romania	1	-	-
Germany	-	-	115
India	-	-	8
Singapore	-	-	9

9 Conclusion

We made a comprehensive research and analysis on the well-known cyberspace search engines such as Shodan, Censys, BinaryEdge, ZoomEye and Fofa. We deeply analyze the items of Supporting internet protocols, Total amounts of detected devices, Device information, Scanning frequency, System architecture, The third party databases, Probes distribution. This paper give an objective evaluation of the detecting abilities and the working principles of the cyberspace search engines by querying official websites, analyzing APIs, and making academic research. We believe this paper will greatly help those who are developing and using cyberspace search engines.

References

1. https://api.shodan.io/shodan/protocols
2. https://support.censys.io/hc/en-us/articles/360038762031-What-does-Censys-scan-
3. https://www.zoomeye.org/doc? The channel = user# d - service
4. https://www.freebuf.com/articles/ics-articles/196647.html
5. https://docs.binaryedge.io/modules/
6. https://cli.shodan.io
7. https://censys.io/ipv4/report? Q = &
8. https://www.zoomeye.org/component
9. https://fofa.so/library
10. https://censys.io/ipv4/help/definitions? Q = &
11. Durumeric, Zakir, et al. "A search engine backed by Internet-wide scanning." Proceedings of the 22ND ACM SIGSAC Conference on Computer and Communications Security.
12. https://www.slideshare.net/balgan/binaryedge-presentationbsides
13. https://greynoise.io/