JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CRAWLING OF DARKWEB FOR SELLING DRUGS ILLEGALLY

[1]Thenisha S, [2]DharaniKumar A B, [3]Arun K, [4]Abirami A, [5]Lakshmanaprakash S
[1, 2, 3,4,5]Department of Information Technology,
[1, 2, 3,4,5]Bannari Amman Institute of Technology, Erode, India

Abstract:
This functionality is designed to create an individual login to the Tor network that can access, analyze and evaluate websites containing drug samples. The overall goal of this project is to create a search engine that allows users to search for websites with illegal content on the TOR network, search, identify and identify hidden services and black markets, improving the current search accuracy of engine results. He creates models on the dark web that make pharmacies illegal. It predicts chemical composition, chemical composition and supplier of origin (India). We offer different technologies like scrapy to build a browser, docker stores the list of servers and virtual machines and uses Apache solr and MongoDB as database. The type of algorithm the browser uses is DFS or Deep Search. However, in this case the interest is in broad search rather than deep search, so the algorithm to use is BFS or broad priority search.
Index Terms – Dark Web crawling, TOR Network, Criminal patterns, Illegal drugs, Darknet, black markets.

I. INTRODUCTION

The dark web is essential for the web that isn't noticeable to web search tools and requires the utilization of an anonymizing program called Tor to be accessed. This dark web will allow people to be involved in illegal activities. It makes the path to buy drugs, guns, credit card numbers and many more things that are not possible on the surface web. The expressions "dark web" and "deep web" are some of the time utilized reciprocally, however they are not something very similar. Dark web alludes to anything on the web that isn't recorded by and in this manner, not accessible via a web search tool like Google. Dark web content incorporates anythingbehind a paywall or requires sign-in qualifications. It additionally incorporates any satisfaction that its proprietors have obstructedweb crawlers from ordering. Clinical records, expense-based content, participation sites, and classified corporate pages are only acouple of instances of what makes up the dark web. Gauges place the size of the profound web at somewhere in the range of 96%and the vast majority of the web. Just a minuscule piece of the web is open through a standard internet browser — for the most partknown as the "clear web". The deep web is a subset of the dark web that is purposefully covered up, requiring a particular program — Tor — to access, as made sense underneath. Nobody truly knows the size of the deep web, however most gauges put it at around 5% of the absolute web. Once more, not everything the deep web is utilized for unlawful purposes regardless of its inauspicious sounding name. Not everything on the dark web is illegal. The Tor network started as a mysterious correspondences channel, it actually fills a significant need in assisting individuals with imparting in conditions that are threatening to free discourse. Many individuals use it in nations where there's listening in or where web access is condemned.

The main objective of this architecture/system is to

- To Estimate the amount of drugs, drug content and vendors of drugs.
- Filter out websites based on shipping from origin/location.
- Develop a system/architecture to precipitate the illegal drug selling websites on darknet.

I. LITERATURE REVIEW

Prominent writers separately offer their viewpoints on the challenges posed by dark web and deep web and illegal activities heldthere. They serve as a launchpad for additional research and demonstrate the importance and relevance of the current work.

The paper proposes the idea of Minimum Executable Pattern (MEP), and afterward presents a MEP age technique and a MEP- based Deep Web adaptive query strategy. The query strategy expands the question interface from single textbox to MEP set, and produces neighbourhood ideal questions by picking a MEP and a keyword vector of the MEP[1].

Internet and network technologies have developed decisively over the most recent twenty years, with rising clients' requests to safeguard their identities and protection. Researchers have created ways to deal with clients' requests, where the greatest piece of the web has shaped the Deep Web. However, as the Deep Web gives the resort to numerous harmless clients who want to protect their security, it likewise turned into the ideal floor for facilitating unlawful exercises, which produced the Dark Web.[2]

The specialists established that there exists a piece of the secret web known as the "Dark Web". The idea of the Dark web charmed scientists to gather data and present inference about the data spread and correspondence occurring on this clouded side of the internet. The introduced work Spy Dark is one such endeavour to gather data from the surface and dark web.[3]

In: AMCIS (2005) Proceedings The writer has explained Building knowledge management system for researching terrorist groups on the Web.[4]

The paper "A Focused Crawler for Dark Web Forums" by Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen (2010) focuses on the development of a system for crawling Dark Web forums. The Dark Web is a part of the internet that is not indexed by traditional search engines and requires specialized software to access. This research is relevant because it addresses the challenge of collecting data from online communities that operate in a hidden and potentially illegal manner.[5]

"Pannu, Mandeep., I. K. H. (2018). Using Dark Web Crawler to Uncover Suspicious and Malicious Websites, in In:International Conference on Applied Human Factors and Ergonomics.", describes a research paper presented at the International Conference on Applied Human Factors and Ergonomics in 2018.[6]

"Sriram Raghavan, H. G.-M. (2001). Crawling the hidden web, In: Proceedings of the 27th VLDB Conference, Roma, Italy.", describes a research paper titled "Crawling the Hidden Web" by Sriram Raghavan and Hector Garcia-Molina. The paper likely explores the concept of the hidden web and techniques for crawling it. The hidden web, also known as the deep web, refers to the vast amount of information on the internet that is not indexed by traditional search engines. This information can reside behind search forms, require logins, or be dynamically generated, making it invisible to standard crawlers.[7

"M. B. D. B. Onur Catakoglu. (2017). Attacks landscape in the dark side of the web, In: Proceedings of the Symposium on Applied Computing, ACM, 2017.", describes a research paper presented at the ACM Symposium on Applied Computing in 2017. The focusof this paper is likely on understanding the types of attacks and attacker behavior prevalent in the dark side of the web**.** The "dark side of the web" often refers to the Deep Web, which encompasses areas not indexed by search engines. This includes password- protected pages, dynamic content, and the Dark Web, a subset of the Deep Web requiring specialized software for access and oftenassociated with illegal or controversial activities.

II. PROPOSED METHODOLOGY

For most of the customers, Google is the main entry point to web browsing. However, the deep web contains pages that Googlecannot crawl. Here are dangerous web anonymous websites, often called secret socities, that control crimes ranging from drugs to hacking and crime. Website URLs on the dark web do nit follow the schedule and often contain random letters and numbers followedby the .onion subdomain. These sites require the TOR program to be taken into account and cannot be accessed through traditionalprograms such as Chrome or Safari. Create a browser that targets the TOR hidden service. By accessing websites containing drug samples, it extracts information such as the seller's username, drug, country of delivery, product description and payment options. These pharmacies facilitate the trade of medicine by using advanced technology to ensure the anonymity of participants and deliverproducts. This study provides an overview of the Indian pharmaceutical industry using data collected through a systematic approach. It will help the authorities understand the most popular drug dealers and products sold by selected countries. It also provides information about the structure and organization of current online pharmaceutical sales. You can access information such as supplierdiversity, industry wide competition, the number of brands each supplier controls, the number of pharmacies it operates, and the products it sells. Analysis of data results in insight into drug sales. Additionally, darknet drug trafficking is difficult to detect basedon digital data alone. It must rely on the combination of physical data in a model to produce results. Confidentiality and integrity areguaranteed on the Tor network. However, on the dark web, the truth is a bit difficult to use because servers are constantly changingand moving. Therefore, digital certificates do not matter much on the Tor network. Some connections may drop because they respond with 3xx (redirect), 4xx (client error), or 5xx (server error) HTTP codes. For 3xx replies, the replaced website will be indexed andthe old website will be discarded. For connections with HTTP codes 4xx and 5xx, the response is simply discarded. The 2xx HTTPcode will be added directly to the file.

The crawler worked on the way to identify the links that are based on the pattern selling drugs. Hidden wiki was the only way to get the bunch of urls that are related to the illegal drug markets. In the dark web there is no sign to get the perfect url and it is not easy to memorize the random 56 characters ending with .onion link. So an easy way to scratch up the crawler was to start with links that are displayed in the hidden wiki. Once the crawler starts with the first link it will crawl through the website and find the internal links and external links of the website. Internal links refers to the site that are related to the same domain (ie. subdomains).External links are the links that redirect to the other links. The crawler crawls till the fixed range set on the frontier to avoid loopingin the dark web and malfunctioning. Scraping takes place in the final part of each url to get the details. An environment needs to be set up to perform the task. The performance of the crawler depends largely on the characteristics of the environment. We had aplan to configure the environment based on the below configuration. 2 GB of RAM was allocated to docker to process the imagesduring the development of the project. 3 GB of RAM were assigned to the virtual machine running the Kali OS in where the Apache Solr search engine is located. The rest of the RAM was divided between the host operating system Windows 10 with its

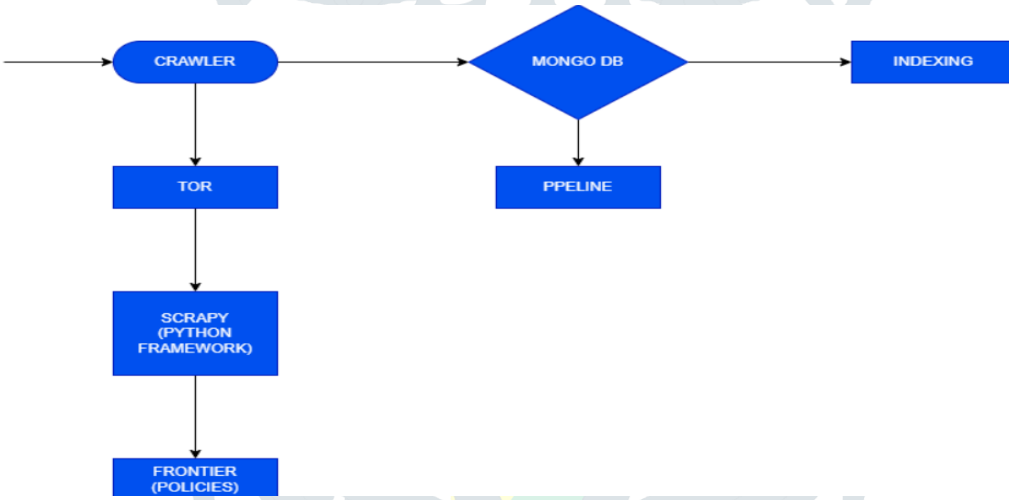
processes and Python for the crawler execution. The memory and the processing capacity can directly affect the performance of the crawler. Since the configurations made in the preparation are made with these capacities in mind.

Table 2.1. Prediction of results

Features	**Results**
Seed URLs	1
Collected Links	171
Discarded Links	23

The first execution was conducted with only one of the most visited seed Urls in the dark web, the crawler was executed for 5 minutes and indexed 171 links. It was inactive for a while; more URLs were collected to serve as seeds. Once the amount of 15 seeds was collected includingthe initial one, it was executed again in which it indexed around 1000 seeds, that is to say, with the triple of seeds it indexed approximately the triple of URLs.

Fig:2.1. Technological stack



Technical Specification:

- AMD Ryzen 7 processor upto 8 cores with 4.7GHz clock speed
- RAM - DDR43400 MHz 12 GB,
- GPU - NVIDIA GeForce GTX 960M (4GB),
- Storage - 1 TB HDD 512 GB SSD
- LAN - 50 Mbps Ethernet

III. IMPLEMENTATION

A. Seed URLs: It was the starting set of URLs. It was taken from the Hidden Wiki website characterized by listing important links from the dark web. It was constantly updated so it's a good starting point for crawling. These dark website links are changed frequently or not available at all. So these URLs must be kept updated for better performance.

B. Crawler: The set of Seed URLs enter the crawler. It will be executed one by one in parallel using threads. In each of these threads crawler takes the URL and sends it to the Privoxy proxy server. It converts the HTTP request to SOCK5 so that the requests enter the TOR network.

C. TOR: The Tor network will respond to that request according to the availability of the server being queried, so it may or may not return the webpage that is required. This response reaches the crawler, which takes this request and redirects it to the spider for processing.

D. Spider: If the response is valid and has content in it, then the spider will analyze the content and collect all the data. This data includes the domain name, URL address, title, content and HTTP status code.

E. Scrapy Link Extractor: Then the spider sends the response to the scrapy link extractor. It extracts every single link on the page and increases the border of the crawler to crawl. These URLs enter the queue for the crawler to analyse them and start the process again.

F. Frontier: It represents the data structure used for storage of URLs eligible for crawling. It makes up the architecture of the crawler. It contains the logic and policies of the crawler.

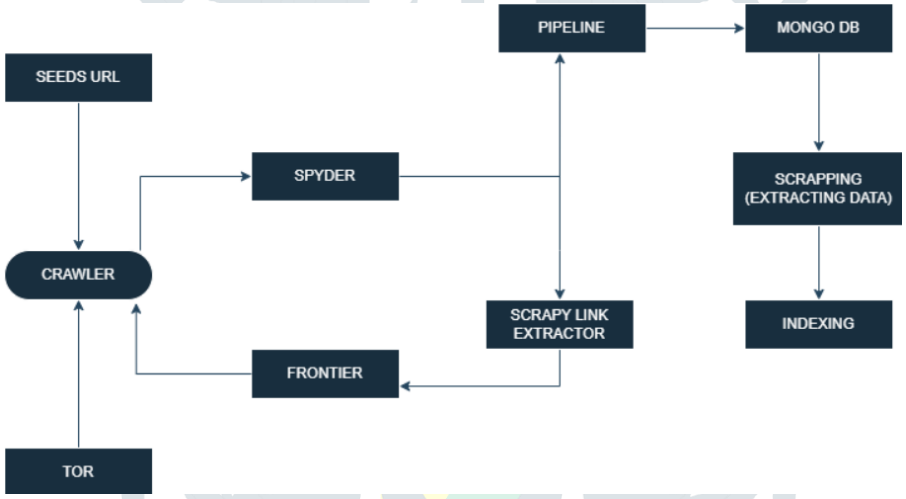
G. Item Pipeline: It opens connections to the database. It analyses whether to store or not store the data. With the help of these

H. MongoDB: The crawled URLs are stored on the MongoDB database in the BSON format. BSON represents the Binary Javascript Object Notation. It can later be converted into CSV format and into many more formats.

I. Scraping: Scrap the data from the crawled URLs stored on the MongoDB one by one. The scraped data contains usernames of the vendors, drugs offered for sale, shipping from the country, description of the products and acceptable payment options.

J. Indexing: Need to collect the data and store it in the manner that it should be served whenever it is needed.

Fig:3.1. Dark web crawling workflow



- It crawls the websites on Tor containing drug selling patterns and extracts data's.
- Spider collect url address, HTTP status code, Domain Name by analyzing content.
- Scrapy link extractor extracts link to crawl deeper.
- Frontier (Data Structure) contains logic and policies of crawler.
- MongoDB to store collected links as JSON file.
- PGP session keys to particular Vendor is found and the location/origin is found.

IV. RESULTS AND DISCUSSION

In this paper we present the architecture of the dark web crawler which searches, distinguishes, and records mysterious administrations, underground markets, and criminal examples. We used python to build up the crawler and the virtual machine with Mongodb as the database. This dull web-cantered crawler ended up being proficient in the pursuit for secret administrations on account of its fundamental component which is being centered. Working on now is the ideal time and results in an extraordinary manner contrasted with an overall crawler. Later on, work we will carry out a savage power framework which will make URLs and analyze them individually. With this technique we will actually want to find site pages that are not recorded in the seeds. Furthermore,we will relocate from a neighborhood server to a cloud server to have the option to distribute the highlights the framework needs with moderately low expenses.

Fig:3. Tor Crawler Developed

The above is the crawler developed and it has five functions available. The first one is to crawl the internal endpoints and the second one is to crawl the internal and the external links and the following are to scrap the web pages, to download the images and the last is to exit from the crawler.

REFERENCES

[1] A. Celestini and S. Guarino, "Plan, Implementation and Test of a Flexible Tor-OrientedWeb Mining Toolkit,"in ACM Worldwide Conference Proceeding Series, Jun.
[2] Sherman, Chris G. P. (2003). The Invisible Web: Uncovering Sources Search Engines Can't See, LibraryTrends, 52, (2) 282-298, 2003.
[3] Barik, K., Abirami, A., Konar, K., Das, S. (2022). Research Perspective on Digital Forensic Tools and Investigation Process. In: Misra, S., Arumugam, C. (eds) Illumination of Artificial Intelligence in Cybersecurity and Forensics. Lecture Notes on Data Engineering and Communications Technologies, vol 109. Springer, Cham.https://doi.org/10.1007/978-3-030-93453-8_4
[4] Abirami, S. (2019). A Complete Study on the Security Aspects of Wireless Sensor Networks. In: Bhattacharyya, S., Hassanien, A., Gupta, D., Khanna, A., Pan, I. (eds) International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 55. Springer, Singapore https://doi.org/10.1007/978-981-13-2324-9_22.
[5] Sriram Raghavan, H. G.-M. (2001). Crawling the hidden web, In: Proceedings of the 27thVLDB Conference, Roma, Italy.
[6] M. B. D. B. Onur Catakoglu. (2017). Attacks landscape in the dark side of the web, In: Proceedings of theSymposium on Applied Computing, ACM, 2017.
[7] Hawkins, B. (2016). Under The Ocean of the Internet - The Deep Web, 15 May 2016. [Online]. Available: https://www.sans.org/reading-room/whitepapers/covert/ocean- internet-deep-web-37012
[8] M. B. D. B. Onur Catakoglu. (2017). Attacks landscape in the dark side of the web, In:Proceedings of the Symposium on Applied Computing, ACM, 2017.
[9] "Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation" Bassel AlKhatib,Randa Basheer Syrian Virtual University, Syriat_balkhatib@svuonline.org randa.s.basheer@gmail.com.
[10] Mundluru, Dhirendranath., X. X. (2008). Experiences in Crawling Deep Web in the Context of LocalSearch, In: 5th Workshop on Geographic Information Retrieval.
[11] Ling Liu, M. T. O. z. (2018). Encyclopedia of database systems, Springer.
[12] SpyDark: Surface and Dark Web Crawler, Ashwini Dalvi; Swapneel Paranjpe; Riddhi Amale; SarveshKurumkar; Faruk Kazi; S.G. Bhirud
[13] Web Crawler for Searching Deep Web Sites, Tejaswini Arun Patil; Santosh Chobe
[14] "Google's Deep-Web Crawl" Jayant Madhavan, David Ko, Łucja Kot , Vignesh Ganapathy, AlexRasmussen , Alon Halevy.
[15] S. Byers, J. Freire, and C. T. Silva. Efficient acquisition of web data through restricted query interfaces. InWWW Posters, 2001
[16] P. G. Ipeirotis and L. Gravano. Distributed Search over the Hidden Web: Hierarchical Database Samplingand Selection. In VLDB, pages 394–405, 2002.
[17] P. Wu, J.-R. Wen, H. Liu, and W.-Y. Ma. Query Selection Techniques for Efficient Crawling of StructuredWeb Sources. In ICDE, 2006.
[18] "Crawling and Mining the Dark Web: A Survey on Existing and New Approaches" MohammedKhalafallah Alshammery, March 2022.
[19] "TOR Gizli Servis Tarayıcılarının Performans Karşılaştırması" December 2019 Bilecik Şeyh EdebaliÜniversitesi Fen Bilimleri Dergisi 6(2) DOI:10.35193/bseufbd.608555.
[20] Priya, R.L., Abirami, A., Desai, N. (2022). Machine Learning-Based Emerging Technologies in the Post Pandemic Scenario. In: Chang, V., Kaur, H., Fong, S.J. (eds) Artificial Intelligence and Machine Learning Methods in COVID-19 and Related Health Diseases. Studies in Computational Intelligence, vol 1023. Springer, Cham. https://doi.org/10.1007/978-3-031-04597-4_3.