International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, September 2015
www.ijates.com
ijates
ISSN 2348 - 7550

GOOGLE'S ONGOING DEEP DRILLING VENTURE

Raman Solanki[1], Akshat Kapoor[2]

[1]*IT, Guru Gobind Singh Indraprastha University, (India)*

[2]*BTech, IPEC UP, (India)*

ABSTRACT

*The **Deep Web**, **Deep Net, Invisible Web** or **Hidden Web** are search terms referring to the content on the World Wide Web that is not indexed by standard search engines. Recent advances of internet over the two decades of more than two billion users. The expansion resulted in developing applications for the cyber world, however there are various applications not accessible. This hidden web over the past years been used as an invariable tool and is not regulated or monitored. Google is giving their full effort to give user necessary access and results but still the future of the deep web is uncertain.*

Keywords: Deep Web, Dark Web, Deep Web Crawling, Google Deep Drilling

I. INTRODUCTION

Deep Web can be delineated as an ambiguous description of the internet not necessarily accessible to search engines. Dark web and the Deep web are used interchangeably however they are not same. Dark web refers to any web page that has been cancelled to hide in the plain sights or reside within a separate but public layer of standard internet. The internet is constructed around web pages that refer other web pages, if you have a destination web page which has no inbound links you have cancelled that page and it cannot be found by users or search engines. Virtual Private Networks (VPN) are another outlook of the crepuscular web that exist within the public internet which require additional software to access.TOR (The Onion Router) is a vivid example. Hidden within the public web is an entire network of different content which can only be accessed using the TOR network. While personal freedom and privacy are admirable goal of the TOR network the ability to travel over the internet with complete anonymity nurtures a platform ripe for what is considered illegal activity in countries like India and China. Dynamic web pages, blocked sites, unlinked sites, private sites, non HTML /-contextual/-scripted contents and limited access networks is not indexed by known search engines like Bing and AOL. The surface web which people use routinely consist of data that search engine can find and then offer a in response to queries, this is only the tip of the iceberg. A traditional search engine sees about 0.3% of the information that is available.[1] Much of the rest is embedded which is called the Deep Web also known as „Hidden Web", „Invisible Web", and the „Undernet". Most of the content located in the deep web exists in the websites that require a search that is not implicitly illicit. Dark web is a very dynamic place; an online forum can be at a specific URL one day and gone the next day. The naming and addressing schemes in the dark web often change this means that the information be harvested two weeks ago is no longer relevant today.

II. DARK WEB

The dark Web is the sector of the deep Web that has been purposely hidden and is inaccessible through standard Web browsers. Dark Web sites serve as a platform for anonymity as essential for Internet users, since they not only provide safety from unauthorized users , but also usually include encryption to prevent monitoring.

A comparatively known source for content that stays on the dark Web is found in the TOR network. The Tor network is a group of operated servers who have volunteered to allows people to upgrade their privacy and security on the Internet that can only be accessed with a special Web browser. First came up as The Onion Routing (TOR) project in 2002 by the US Naval Research Laboratory, it was a method for online communication anonymously. Another network, I2P, provides many of the same features that Tor does. Though, I2P was designed to make a network within the Internet, with traffic staying contained in its borders. Tor provides superior anonymous access to the open Internet and I2P provides a more robust and reliable "network within the network" [2][4][10].

2.1 How to Access Dark Web?

- One widely used is TOR browser is one way to access dark web sites as they are in .onion extension that cannot be accessed by normal browsers so TOR Browser is used for browsing the web to withhold some information about your computer's configuration.
- Tor2web created by Aaron Swartz and Virgil Griffith. It is bridge between public internet and untraceable sites.

  However it won‟t be anonymous in the way they would be if they used Tor .[7]

2.2. How does TOR Network Works ?

Tor protects you against a familiar form of Internet surveillance that is traffic analysis. Traffic analysis can be used to deduce who is interacting to whom over a public network. As the source and destination of your Internet traffic is known this allows others to track your behavior and interests. Tor distributes your transactions over several places on the Internet that reduces risks of both simple and complex traffic analysis, so no single point can link you to your destination. This is similar to adopting a swirly, hard-to-follow route in order to get rid of somebody who is tailing you and then systematically erasing your footprints. Rather than taking a direct route from source‟s database to destination, data packets on the TOR network take‟s a random pathway through many relays that cover‟s your tracks so that no observer at any single point can tell from where the data came or where it's going to create a private network pathway. In TOR, the user's software incrementally produces a circuit of encrypted connections by relaying on the network. The circuit is elongated one bounce at a time, and each relay along the way knows only which relay gave it data and which relay it is giving data to. No individual relay can ever know the complete path that a data packet has withdrawn. The client accommodates a separate set of encryption keys for each bounce along the circuit to ensure that each bounce can't trace these connections as they pass through.[6]

Random pathway network User

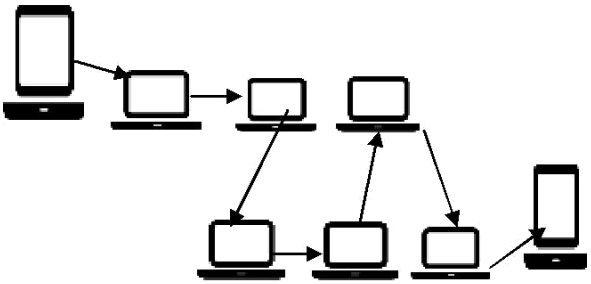

Figure 1- Working of TOR Network

III. DEEP WEB

The deep web is that part of the internet not accessible to link-crawling search engines like Google; the only way the user can access this portion of the internet by typing a directed query into a web search form thereby retrieving the content from the database that is not associated. The only way to access the Deep Web is by conducting a search that is within a particular website. Surface web search engines can lead you to website that has unstructured deep web contents. For instance if one need to search for a published article, there are certain papers that are only accessible if one has access to the Google Scholar web pages, which lead to the database connected to it, there is no other way; a person or an organization can access the database if they don‟t have the correct en-route to it. If you are searching for some government grants the search engine, will direct you to the website www.grats.gov, researchers can search thousands of grants by searching the database via the website search box. In this surface search engine lead users to the deep web websites where the directed query to the search box brings back Deep web content not found via the search engine.[8][9]

Figure 2- Deep Web Location

3.1 Deep Web Search Engines

Search engines such as Google indexes over a trillion pages on the World Wide Web, but there are information on the web that common search engines are unreachable. Most of them are in databases that needed to be searched directly from the particular website. A small pocket of the deep web is filled with hyper-secret communities who gather there in order to escape identification from authorities. To discover content on the Web, search engines use web crawlers that uses a technique to follow hyperlinks. This technique is ideal for discovering resources on the surface Web but is often feckless in finding deep Web resources. For example, these crawlers do not attempt to find dynamic pages that are the result of database queries because of the infinite number of queries that are possible. It has been acclaimed that this can partially be overcome by providing links

to query results, but this could unintentionally inflate the popularity for a member of the deep Web example is the Page Rank. In 2005, Yahoo made a small part of the deep web searchable by the release of Yahoo Subscriptions. This search engine searches through a few subscription-only web sites. Some search tools such as Pipl are purposely designed to retrieve information from the deep web their crawlers are set to identify and interact with searchable databases that aim to provide access to deep Web content. Deep web harvesting service provided by Bright Planet .They harvest each and every big data from deep web for the client . It extracts every single word every time it accesses a web page. Plus, the Deep Web Harvester stores every single page harvested as a separate. [7][4][9]

3.2 Deep Web Crawling

Crawling the deep web automatically has been a major aim for some researchers these days. Raghavan and Garcia-Molina presented an architectural model in 2001 for a Deep Web crawler that used key terms provided by users or collected from the query interfaces to query a Web form and crawl into the deep Web resources. Ntoulas et al. created a hidden-Web crawler in 2005 that automatically generated meaningful queries to issue against search forms. Even though crawler generated favourable results, but the problem is far from being solved. [5][9]

IV. GOOGLE'S DEEP DRILLING

Since a huge amount of useful information and data reside in the deep web. Google had to have its ways to crawl in deep web in order to index the web pages. Since 1996 Google is remodelling each year and broaching new techniques and ways to crawl in the deep web and indexing sites needed. Though Web crawlers are the central part of the search engines so their design and major architecture in kept business secret. Whenever the Crawler design is published there is always a lack of information that prevents reproducing the work from others because there is always of spamdexing. Increase in spamdexing in the mid-1990s made the leading search engines of the time less useful. Spamdexing is using unethical method to make web pages rank higher in search engine results than they otherwise would is commonly referred to in the SEO (Search Engine Optimization). [11]

4.1 Googlebot

Since Google is made by using web crawler to index sites that are linked to one and another in order to give you better search results.

4.1.1. Continuous chain of events causing Googlebot existence

Early on, Google is primarily streamlining what was BackRub which was a python based crawler. Google team converted the crawler to C++ and started working on a distributed crawling system to overcome the massive architecture challenges related to web-scale crawling and indexing. 2003 Google‟s 1st update came out "Boston" Announced at SES Boston. Google first, aimed at a major monthly update, so the first few updates were a medley of algorithm changes and major index refreshes. As updates became more frequent, the monthly idea quickly died. [moz.com]April of 2003

"Cassandra" update arrived that had major changes to the algorithm to detect hidden links and text. From 2006 to 2008 years represent feature enhancement updates. Google spends quality time renovating how we interact with search. They got a lot of data, which they shared to make the user experience richer and more useful. 2006 is one of the first appearances of Googlebot 2.1 with its Mozilla5.0 Compatible user-agent came up. 2010 saw the full roll out of new update name "Caffeine". Caffeine not only boosted Google's raw speed, but also enhanced crawling and indexation much more firmly that resulted in a 50% fresher index. This all updated enhanced Google crawling techniques and make Googlebot what it is now.[12]

4.1.2. Understanding Googlebot

"Googlebot" is that web crawler used by the Google which collects info and data from sites to built a searchable index for search engine. "Spider" is the other name given for this. Googlebot uses an algorithmic process where computer programs induce which sites to crawl, how often, and how many pages to withdraw from each site. Googlebot's crawl process begins with a list of webpage URLs, generated from previous crawl processes and augmented along with Sitemap data which webmasters provide. As Googlebot visits each and every of these websites it detects links on each page and include them to its list of pages to crawl. New sites, changes to existing sites, and dead links are noted and used to update the Google index. If website owner downs not want particular part of his site to be crawled by the crawler he can include "robot.txt" file to block access to files and directories on your server. A robot.txt is a file that uses

"Robot Exclution Standard Protocol ", it is a small set of commands that are used to indicate access to site by section and by specific web crawlers. [13][14]

V. SITEMAP PROTOCOL

On 16 November 2006 first joint and open initiative held to improve web crawling in between Google, Yahoo and Microsoft in which they announced support for sitemap 0.90 which is a free and easy way for webmasters to notify each search engines and be indexed more comprehensively and efficiently. This gives better representation in search indices. [15]

5.1 Understanding Sitemap

A Site map is a XML file where web pages of a site are been listed to inform Google and other search engines about the organization of that site content. Search engine web crawlers like Googlebot read this file to more intelligently crawl your site. Sitemap can provide valuable metadata that is information such as, when the page was last updated, how often the web page is changed along with the importance of the page relative to other URLs in the site. Sitemap is made in XML and is submitted in Google Search Console. Any company that that manages dynamic contents or have large number of benefit with Sitemap Protocol. If a site contains a deep web page that is not indexed or crawled by crawler can give sitemap to the search engine in order to indexed and searchable by users.[16][15][17]

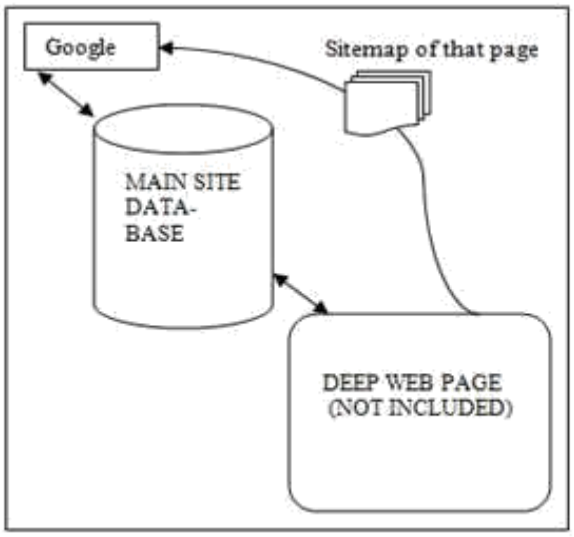

Figure 3- Sitemap Protocol Explained

VI. GOOGLE'S MOD_OAI

Google started encouraging mod_oai projects to harvest better. Mod_oai is an Apache module that allows web crawlers to discover new, updated, and deleted web resources from a web server efficiently by using OAI-PMH, protocol which is used in the large scale by the digital libraries community. mod_oai also allows harvesters to obtain "archive-ready" resources from a web server. In 2005 Google used OAI-PMH protocol to harvest information from the National Library of Australia (NLA) Digital Object Repository. [18][19]

6.1 Understanding Mod_oai

Mod_oai an Apache module that responds to OAI-PMH requests on behalf of a web server. If Apache and mod_oai are installed at for instance http://www.xyd.com/, then the base URL for OAI-PMH requests is http://www.xyd.com/mod_oai. mod_oai discloses the files on an Apache web server as an OAI-PMH archive. Where as OAI-PMH stands for Open Archives Initiative Protocol for Metadata Harvesting. It was built to simplify the process of gathering information or metadata from digital repositories. [20][21]

VII. GOOGLE LIMITS ITSELF

Vertical search engines like CloserLooksearch, Alacra are some specialty engines which crawl subject category or vertical. These sites can search deep dynamic or password protected sites. Where as Google is said to be horizontal search engine but, Google intentionally limits itself. As vertical search can be useful in particular topic search like searching particular database but horizontal search expands its search variety to give user a better, appropriate and wide search quality. The web pages Google surface include listings of results and these pages are inserted into their web index. Spam sites can also be resulted in vertical sites but not by Google. [9]

VIII. CONCLUSION

Since a huge amount of useful information and data reside in the deep web; its engine has begun exploring alternate methods to crawl the deep web. Google site map protocols and mod oai are the mechanism that allows the search engine and other search parties to discover the deep web resources on particular web servers. Both mechanisms allow the web server to advertise the URL"s that are accessible on them thereby allowing automatic discovery of the resources that are not linked directly to the surface web.

Google may have come long way in crawling but still there are some parts which are out of their reach like dark web. For deeper information we have to use vertical search engines, so Google can improve their interface in order to search as subject category in their options. In the near future, the deep web will be explored to all its content and further deep web will develop.

REFERENCES

[1] UNDERSTANDING THE DEEP WEB IN 10 MINUTES, Steve Pederson, CEO,Brightplanet.

[2] BELOW THW SURFACE EXPLORING THE DEEP WEB, Dr. Vincenzo Ciancaglini, Dr. Marco Balduzzi, Robert McArdle, and Martin Rösler Forward-Looking Threat Research Team

[3] THE IMPACT OF THE DARK WEB ON INTERNET GOVERNANCE AND CYBER SECURITY,Michael Chertoff and Tobby Simon February 2015.

[4] Bergman, Michael K. 2001. "White Paper: The Deep Web: Surfacing Hidden Value." http://quod.lib.umich edu/j/jep/3336451.0007.104?view=text;rgn=main.

[5] EFFICIENT, AUTOMATIC WEB RESOURCE HARVESTING Michael L. Nelson, Joan A. Smith and Ignacio Garcia del Campo.

[6] TOR:OVERVIEW,https://www.torproject.org/about/overview.html.en.

[7] NEW SERVICE MAKES TOR ANONYMIZED CONTENT AVAILABLE TO ALL, http://www.wired.com/2008/12/tor-anonymized/

[8] INVISIBLE WEB by closurelooksearch.com.

[9] INVISIBLEWEB/DEEPWEB,http://www.voodish.co.uk/articles/invisible-web-deep-web/.

[10] THE INVISIBLE INTERNET PROJECT(I2P), https://geti2p.net/en/about/intro.

[11] DIGITAL MARKETING HANDBOOK
https://books.google.co.in/books?id=L4wr9mFq7nkC&pg=PA80&lpg=PA80&dq=when+google+used+mod-oai&source=bl&ots=qw0zEBoX7z&sig=KAIjLSoIcE6VRLfxDtlCe5OShOs&hl=en&sa=X&ved=0CDoQ6AE wBGoVC hMIr_GY15z7xwIViReUCh0-2Q-i#v=onepage&q=when%20google%20used%20mod-oai&f=false.

[12] GOOGLE ALGORITH CHANGE HISTORY, https://moz.com/google-algorithm-change.

[13] https://support.google.com/webmasters/answer/1061943?hl=en.

[14] https://support.google.com/webmasters/answer/182072?hl=en

[15] MAJOR SEARCH ENGINES UNITE TO SUPPORT A COMMON MECHANISM FOR WEBSITE SUBMISSION, http://googlepress.blogspot.in/2006/11/major-search-engines-unite-to-support_16.html.

[16] https://support.google.com/webmasters/answer/156184?hl=en. [17]WHAT IS SITEMAP?, http://www.sitemaps.org/.

[18] http://dbpedia.org/page/Mod_oai.

[19] MOD_OAI PROJECT AIMS AT OPTIMIZING WEB CRAWLING, by Michael nelson, releasing date 21 April 2004.

[20] MOD_OAI: AN APACHE MODULE FOR METADATA HARVESTING, by Michael L. Nelson, Herbert Van de Sompel, Xiaoming Liu, Terry L. Harrison, Nathan McFarland.

[21] http://www.modoai.org/.