The Deep Web: Determining the Surface Hidden Value

Shivanshu Gautam, Dr Ankur Goyal

Abstract— **The field of information technology has overridden every knowledge field and almost all the technology available in present. Huge amount of data is available on the web/internet that can be read effortlessly by anyone and wherever across the world over the internet. The internet has amplified the craving of the masses to rely on it, to the degree that the need can be said to be nearly comprehensive. The facts on World Wide Web is aggregating spirally causing into large amount of information to be thrown out of the surface. The extraction of the relevant but thrown out data is becoming challenging chore day by day as the traditional network crawlers focus only on the information that is available on the surface of this web network, wherein the contents in deep web is sized as, some multiple times as that of the surface web but deep web is expanding at exponential rate. The Deep web data could be fetched using the query interfaces. The Hidden Web Extractor generically called "the HWE", which assists in finding the relevant information and harnessing the info out of the Deep Web repositories and hence the query interface is the only instrument for tracking the most relevant data out from the Hidden part of a website. The Hidden Web Extractor has supported the user in retrieving infinite number of website pages through the significant outlines of queries.**

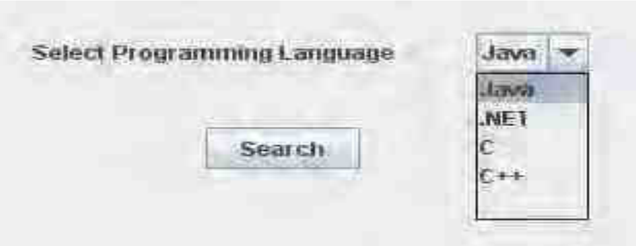
Index Terms— **The Deep Web, Dark Web, Relay, Tor, legal activity, browser, research, Surface Web**

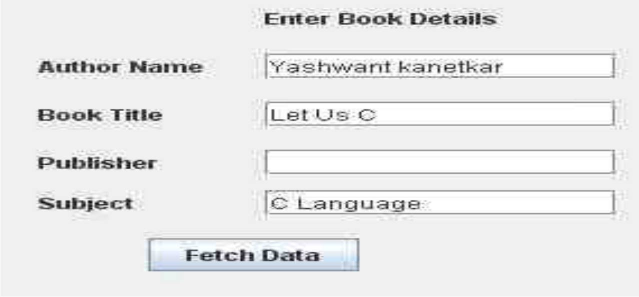
I. Introduction

The Deep web page was started in year 1994 known as Hidden Web which later was renamed as Deep Web in 2001. Also the network created as Tor, famously known as "Deep Web", was established and supported by the U.S. Naval Research Laboratory with the hope that it shall help to protect and encrypt the data under government communications. It would create confident and protected communications so as to escape censorship as a way to assurethe law under free speech. For example, The Deep Web facilitatedorganize the Arab Spring Protests [1]. But evidently unlikely any alternative tool, the impact caused shall get change based on user to user. Now aided by the Tor Project, expecting from year 2004 to 2005 it was supported by the Electronic Frontier Foundation. The Deep Web projects are supported financially by different government institutions across the world, including the United States and India to this day. Our most commonly known search engines falls short to citeloads of this immensely deep infowhich existsover the internet. Of all these reasons only the Deep Web is commonly compared with an iceberg [2].Theinfo above the level surface is taken as the part of the "searchable Internet," and everything which is beyond that layer is considered as the Deep Web. Information bundled under the piles of data, is not easily visible and becomes difficult to get to anycontent that can be indexed by any search engines like Google and Yahoo. This characterisationthereforeconsist of Unlinked web content, dynamic web pages, blocked sites and URL's, unlinked sites, private sites, non HTML/-contextual/-scripted content, and limited-access networks [3].

SERVERS
THE CLOUD
THE WEB SEA
Google
DEEP WEB
CONTINENTAL CONTENTS
GOVERNMENT
WIKILEAKS
UNDETECTED
ILLEGAL PORN

However the concept of Deep Web is little understood, but it is quite simple. We need to consider it with respect to the search engines. To fetch results for your queries search engines like, Google, Yahoo and Microsoft's Bing constantly index pages. This process is done by crawling the Web's pages like connected threads by following the references between the pages and websites. Which only helps them collect static pages, which are often useless. All the pages behind private grids or standalone pages that connect to nothing are not captured by search engines like Google. These are all part of the Deep Web. By typing a focused query into a web search form, a consumer can reach this portion of the internet and he/she can salvagethe content from a databank that is not interconnected.





Enter Book Details

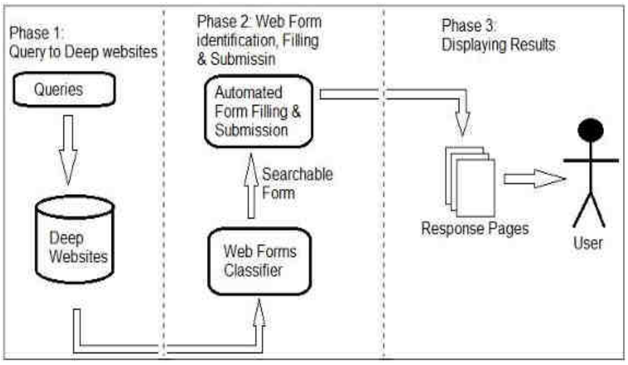
Author Name	Yashwant kanetkar
Book Title	Let Us C
Publisher	
Subject	C Language

Manuscript received August 15, 2019

Shivanshu Gautam, M.Tech(CSE) Scholar, Yagyavalkya Institute of Technology, Sitapura, Jaipur
Dr Ankur Goyal, Yagyavalkya Institute of Technology, Sitapura, Jaipur

II. INFRASTRUCTURE OF THE DEEP WEB

The structural design of deep web includes of the subsequent major modules as shown in the figure, below anexplanation of each component is presented:



1. UX INTERFACE:

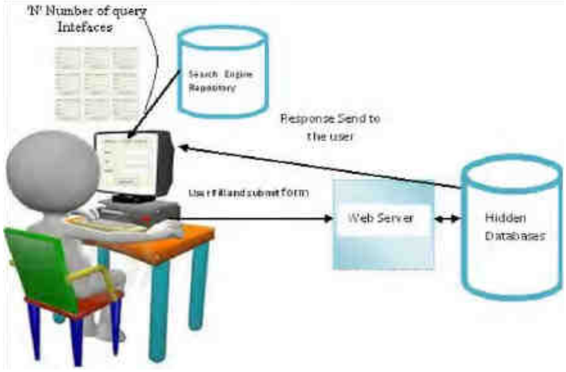
The user experience interface transfers queries by providing aneasy interface to the end-user and then provides the query to the deep web resources thatare related to user specific queries.

2. FORM CONTENTCLASSIFIER:

It verifies whether the identified web form existing in a website is searchable or not. If yes, then the fields in that specific form are acknowledged.

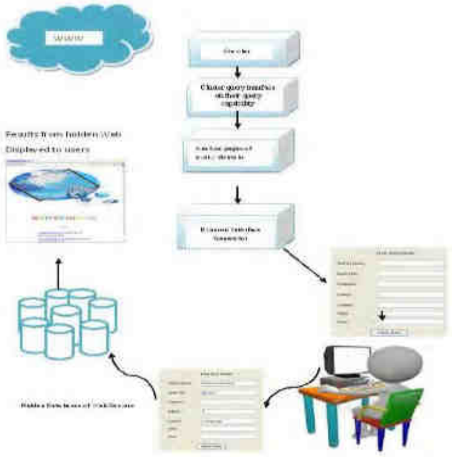
3. AUTO FILLING AND SUBMISSION:

The resultant from the Form Classifier as the searchable form field are recognised and are filled with user provided query,then submitted for response automatically.



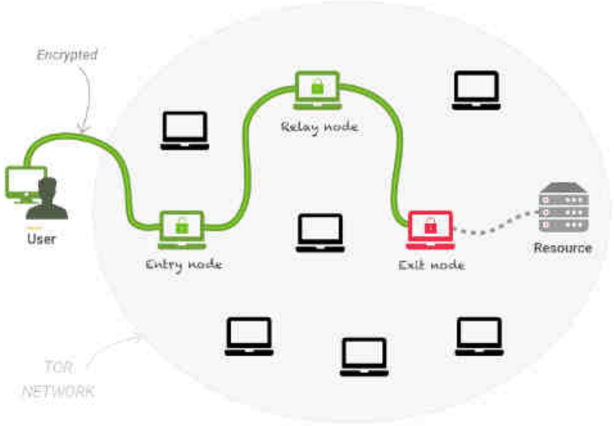
4. RESPONSE RESULTS:

The call-back to the query from the user which was searched by them areshown by submitting the webformtherefore providing the returned response



III. WHAT IS TOR?

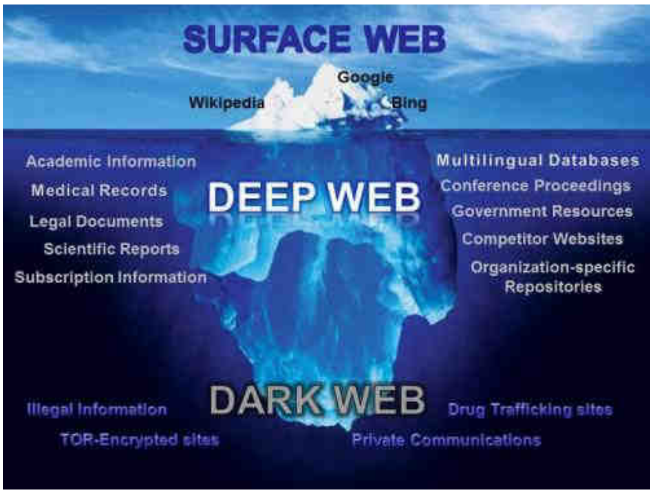
Tor is a networking protocol over the internetdeveloped to hide the information transfer across. Establishing your privacy and security by covering your webmail, browser history, social media engagement posts and all other online activity. If you are using Tor's software, no-one shall be able to tell which location you're in by seeding your IP address, which can be very useful for defence, activists, journalists, business-people and more.



The address layers is encrypted to hide and secure data packets which are passedfrom the Tor network [4]from the user system.

IV. SURFACE WEB VERSUS DEEP WEB:

The Surface Web and the Deep Web are of equal importance. As surface web represent the latest information that is over the web and is in trend or satisfies most of the users. Theshare of the Internet that all standard search engines can catalogue and standard web browsers can access devoid of the need for any kind of special solution and/or configurations is called as Surface Web. It is also acknowledged as “clearnet.”



SURFACE WEB:-

- Entries are statically generated
- Web Crawled Content
- Readily navigable through any browser or search engine unlike the Deep Web, which involvesusage of special search engines, browsers, and proxies.

DEEP WEB:-

- Entries are dynamically generated
- Unlinked Content
- Contextual Web
- Private Web
- Scripted Content
- Non-HTML content
- Limited Access Content

V. TYPES OF INVISIBILITY

Anenormous size of datais available on the internet but cannot be reached through the webreadily for being not indexed properly. The Hidden web or Invisible content on World Wide Web (WWW) are categorized in accordance to their invisibility and categorised in four sections as under:

- The Opaque Web
- The Secret Web
- The Proprietary Web
- The Proper Invisible Web

A.OPAQUE WEB

The Opaque Web is the system of web, which can be indexed by designated web crawlers but are not listed by general purpose search engines or standard Web Crawler due to following reasons:

- Depth of Crawl
- Disconnected URL's
- Frequency of Crawling

B.SECRET WEB

The Undisclosed Web pages are technically listed but not indexed in manner to be openedby general search engines as the Web Adminconstructs the barriers. Web Crawler is not able to catalogue these pages because of the following reasons:

- Meta Tag- No index
- Robots Exclusion Protocol
- Authentication
- Authorization:The web pages may be protected by Password.

C.THE PROPRIETARY WEB

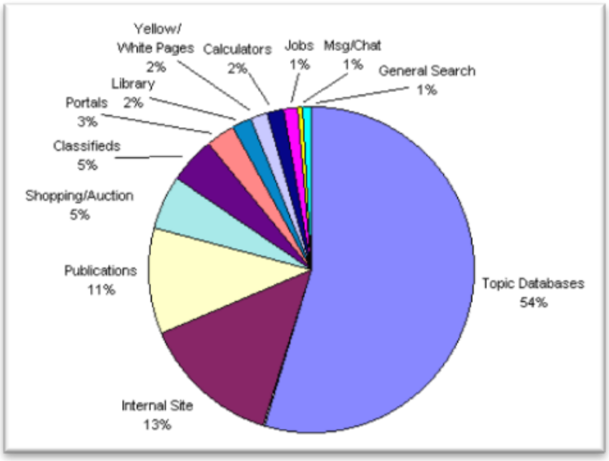
The Proprietary Web has anowner and is controlled like a possessions. Any user eager to view Proprietary Web has to take authorisation from the possessor by first registering. Only the registered ogler can have the view of such pages. The process of registration may be on cost or free of cost as per the owner's desire.

D.PROPER INVISIBLE WEB

Web Crawler may also not be able to index the Proper Invisible Web pages because of format of files which cannot be handled by today's search engines. Ex:

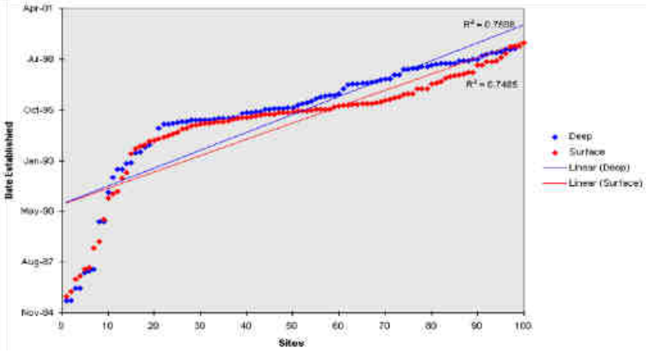
- Audio
- Video
- Compressed Content
- Executable
- Any other application like (office, specific tools extension etc.)

VI. DEEP WEB STATISTICS



Some facts that show the drastic difference between the Surface Web and the Deep Web:-

- Deep web is estimated to be of 7500 terabytes of information as compared to a mere 19 terabytes of information in the surface Web.[9]
- The deep Web consistof almost 550 billion documents Vs.estimated 1 billion of the surface Web.[9]
- Approx. 2, 00,000 deep Web sites presently exist.
- About 750TB of data is contained by merely 60 websites of the deep web collectively — enough to bypass the content of the surface Web forty times.[9]
- On an average, deep websites receive fifty percent more monthly traffic verses surface websites; however, the typical deep website is not well known to the Internet-searching people.
- In comparison to the Surface Web, total content of deep web is approx. 1,000 to 2,000 times bigger.



VII. LEVELS OF DEEP WEB

7.1 THE COMMON WEB

- The web that the vast majority of internet users are accustomed to.
- Examples like Facebook and informational websites like Wikipedia and other general websites, etc.

7.2 THE SURFACE WEB

- Websites from this level is accessible through normal means, but contains "darker" content, such as Reddit and 4chan.

7.3 THE BERGIE WEB

- The websites that are blocked in some nations but exists on the Surface Web,can only accessible through illegal means in these countries by proxy or other means.

7.4 THE DEEP WEB

- The data from the Government/Business/Collegiate Research, that are now not part of the Surface Web and which is indexed by standard search engines are claimed to be the part of Deep Web.

7.5 THE CHARTER INTERNET

- It consists of Prohibited obscene content like CP, Gore, bounty hunters, outlaw game searching, etc. Most of this networkis used to market crypto-currencies, Arms, Drug, Rare Animal Trafficking.

7.6 MARIANAS INTERNET

- It is exceptionallyhard to access it's the safest part of the net owed to non-public users.

CONCLUSION

Though the deep web is infamous for malevolentevents, still we may employ it for productivetenacities. It can also be used for confidentialcommuniqué and interchange of infoamongst intelligence organisations, such as government and defence establishment who need to escape safekeeping from external agencies for concealedactions. Websites like Facebook have engaged an initiative to pawn online expurgation by founding a URL on Tor (facebookcorewwwi.onion).Also, NASA joined hands with Defence Advance Research Projects Agency (DARPA) by making a web browser to make purpose of deep web. If more agencies and computer establishments join their hands to keep a control on the corruptactions on the deep web and proceeds a vigorousedgeof utilization with a positive intention, then only we shall be gifted to use the deep web for the positivetenacities for which it was originallyformed.

As far as information from the static web pages is concerned, with rise in internet users the demand for information shall be increasing day by day. More and more information shall be pushed into deep web with ever increasing posting on the internet and this shall further increase the importance of accessing the relevant information from the deep web. The properly tagging of the information prior to its posting on websites shall help in accessing the information with ease to certain extent but more researches more easy and cost effective extraction of information from the 'Deep Web'.

REFERENCES

[1] http://whitepapers.virtualprivatelibrary.net/DeepWeb.pdf

[2] A Review Paper on Deep Web Data Extraction using WordNet

[3] https://www.technologyreview.com/s/532256/dark-web-version-of-facebook-shows-a-new-way-to-secure-the-web

[4] https://www.vpnmentor.com/blog/tor-browser-work-relate-using-vpn/

[5] http://web.mit.edu/gtmarx/www/anon.html

[6] The Ultimate Guide to the Invisible Web, 2013

[7] http://quod.lib.umich.edu/cgi/t/text/idx/j/jep/3336451.0007.104/--white-paper-the-deep-websurfacing-hiddenvalue?rgn=main;view=fulltext

[8] Below the Surface: Exploring the Deep web (A Trend Labs Research Paper)

[9] https://business-reporter.co.uk/2015/06/22/5-predictions-for-the-future-of-the-deep-web