IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Insights of Deep Web : What Google can't search

Asha Yadav
* Department of Computer Science, Motilal Nehru College,Delhi University,India
yadav.asha26@gmail.com

Abstract

This paper deals with the study of Deep Web that is a hidden unexplored part of internet. Further the characterstics and challenges of searching data in deep web are discussed along with the methods to access such information over the internet. Finally , a comparison is made of various floating term realted to deep web.

Keywords: Deep Web, Dark internet, Hidden Web.

Introduction

We are living in an era of information technology ,thus, the most important and indispensible part of this era is Information undoubtedly. Well by the general definition information is a processed form of data and if we talk about the means to access the information the very first striking word comes through is search . We all know that in today's world internet is one stop destination to all your information cravings . Search engine help us to locate this information from various available sources online. But are they capable of finding everything out there or is there any part of internet resource that is hidden from these conventional search engines. This is what we are going to deal with in this paper.

Searching on internet is more or less like dragging a net across the surface of the ocean.[1] Although many things could be found but still major portion of information remains unexplored reason being the inability of link-crawling spider based search engines to dig deep down.

Definition : Deep Web

"Deep Web" can be simply defined on anything that is available on internet but not necessarily accessible via the tradional search process as the traditional search engines create their indices by spidering or crawling surface Web pages i.e. words in title ,subtitle or meta tags etc . To be discovered, the page must be static and linked to other pages. Traditional search engines can not "see" or retrieve content in the deep Web — those pages do not exist until they are created dynamically as the result of a specific search. Because traditional search engine crawlers cann't probe beneath the surface, the deep Web has heretofore been hidden.

Characterstic & challenges of deep web

The distinguishing characterstics of deep web content not only makes it difficult to locate but also poses a challenge to the search engines to retrieve and display data as per user demand . Michael Bergman's pioneering 2001 study, The Deep Web: Surfacing Hidden Value, estimated that it accounted for 7,500TB of data at a time when search engines could index only 19.[1] Various characterstics of deep net are listed below :

Unstructured

The data in deep web consist of mainly the web databases. Such databases needs tailor-made queries specialy to fetch the user required but since these tailored querries are one at a time its hard to design specific querry for each database as they do not have same structure infact many of them are completely unstructured.

Dynamic

Search engines basically fail to access the content which is generated dynamically over the net. As these pages are generated "on the fly" and they are not stored on web server. Hence indexing such web pages require the search engines to imitate the same action to generate them.

Unlinked

There are many sites that are unlinked and hence undiscoverable. They are almost impossible to locate and indexed by search engines.

Limited and private content

Such content is not indexable by the web crawlers as it require authentication for access and its unlikely for web crawler to generate complex site-specific actions required to gain access to the content.

Importance and Uses

The surface web just consist of 1% of the whole internet rest 99% is hidden in deep web. Bright planet estimates size of the invisible, or deep, web as being **500 times bigger** than the searchable, or surface, Web. Considering that Google alone covers around 8 billion pages.

The deep web data can be used in almost every field the fact being that it provides much more accurate and dwelled information as compared to the surface web.

Mainly the deep web can be used to access the archived source of otherwise visible site. The deep Web is an endless repository for a mind-reeling amount of information. There are engineering databases, financial information of all kinds, medical papers, pictures etc

Data extraction from deep web

As of April 2013, over 144 million unique domains exist online with an estimated 100,000 new domains added daily.

Harvesting

The first step to collect the unstructured , unformatted textual data from various websites as well as their archives. It automates the process of issuing multiple site specific queries used in search boxes of the site and collecting the result from those queries.

Deep Web harvests perform directed queries and harvest only the relevant results. Link following navigates to every single page it can find via hyperlinks, determines if the content is relevant, and then harvests the content if it is deemed relevant.

Normalizing and Enrichment

The data obtained after harvesting is not likely to be in same format as the various forms of data differs from site to site. It may be pdf, text, image, HTML, xls etc. To compare and store these documents, all relevant text needs to be extracted from the harvested data types and stored in one uniform database.

Even in the textual form of data the difference comes in the form of various character encoding, therefore all this data needs to be transcribed in the one specific format so that it can be made presentable.

Then the data is enriched by providing additional structure with metadata, which is generated by an additional extraction on harvested text.

Analytics

Once the data is collected and formatted it can be analysed further by the user using the analysis tools provided by the website. It also incorporates faceted searching that is the ability to break down results of a search into specific categories to make it easy for the user to do categorized search.

Access Techniques

`The deep web refers to all parts of the internet which cannot be indexed by search engines, and so can't be found using Google searches. It can be seen by directly typing the manual querry in the search box of the website. Hence, specialized search engines or browsers are used to dive into the endless information sea of the deep web.

Turbo10

Turbo10 is a commercial search engine designed to access deep Web resources in order to produce search results (http://www.turbo10.com/). The idea is that there are enormous amounts of topic-specific querying interfaces in the Web. These interfaces each produce their own search results when queried.

TOR

The Onion Router is one of the famous browser for deep net. It is a free software to maintain anonymity over the internet It directs Internet traffic through a free, worldwide, volunteer network consisting of more than five thousand relays to conceal a user's location or usage from anyone conducting network surveillance or traffic analysis.

Freenet

The Freenet software and data storage concept was originally developed by Ian Clarke but its under constant development mode since then. So, basically freenet is a free software that helps to share files, browse and publish free website and chat on forum anonymously and without any censorship.Freenet is decentralised to make it less vulnerable to attack.

INFOMINE

A virtual library of Internet resources relevant to university students and faculty. Built by librarians from the University of California, California State University, the University of Detroit-Mercy, and Wake Forest University.

The WWW Virtual Library

One of the highest quality material can be found over here it was Started by none other than Tim Berners-Lee, .

Some similar terms

Since the term deep web becomes so vast that user gets confused between surface , deep and dark web. So, here in this section we deal with the major distinguishing characterstics of all three of them.

Surface net

The conventional web or the visible part of the web that can be easily found via the traditional search approach . It is also known as clearnet or indexable web.

- It mainly consist of linked HTML pages.
- The major content is static and can be indexed.
- They are identified by the spiders or web crawlers via metatags.
- Common search engines are google, yahoo, msn, ask jeeves etc.
- They contain only 1% of the entire web content.

Deep Web

The term coined by Mike Bergman founder of Bright Planet. It consist of that data on the net which is although available for access but cannot be searched using simple web crawling serach mechanism. It requires intense query processing and optimization to access such contents even the resources of such content.

- It mainly consist of archived data bases or dynamic pages.
- The data can be accessed via special automated or manual querying process.
- The pages may or may not be linked.
- The urls are also hidden from the traditional search engines.
- The size is so huge almost 99% of the web content.
- Needs the harvesting of unstructured big data.
- The common search softwares are TOR , freenet,deep peep , intute etc.

Dark web

It is a small portion of deep web that is intentionally hidden .It is only accessible on the TOR network which hides the identity of the browser and the host. The websites here have .onion extension. It is generally linked to the illegal activities over the internet.

Darknet

A darknet is any network where connections are made only between trusted peers using non-standard protocols and ports or using onion routing. Thus Dark Web is a part of the Darknet. However, it is important to point out that the Dark Web and the Darknet are not synonymous. Many other services can run on the Darknet, such as email, IRC, etc. The Dark Web is just one of these services, contributing a subset of traffic over the Darknet.

Dark Internet

It is unrelated term to the above ,yet found quite synonymous.. It refers to the unreachable network hosts on the Internet. They could be unreachable because a machine is turned off, or a network cable is damaged, or even because routing tables have become corrupted somewhere. Nobody, not even regular Internet users, can reach them.
Hence the Dark Internet is constantly changing; as machines get taken offline and put back online, but by the time they form are offline, they are part of the Dark Internet.

Conclusion

Deep net is a vast repository of content which is very useful for the wide and uncovered information about any subject. Although it is not easily accessible still it is not completely hidden.
In this paper , we have got to know some basic but important information regarding the access of deep net. Since it is uncommon to general internet users it is generally confused to be illegal but its not anyhow.
Hence we conclude that deep net has got potential to answer the information cravings of the users if used properly and its resources need to be tapped to full extent.

Refernces

1. Bergman, Michael K (July 2000). *The Deep Web: Surfacing Hidden Value*. BrightPlanet LLC.
2. *BrightPlanet. "How Does Data from the Deep Web go from Results to Actionable Intelligence?" Jan. 31, 2013. (Dec. 6, 2013) http://www.brightplanet.com/2013/01/how-does-data-from-the-deep-web-go-from-results-to-actionable-intelligence/*
3. *Here and Now. "The Deep Web: Where Google Won't Take You." WBUR.org. Nov. 8, 2013. (Dec. 6, 2013) http://hereandnow.wbur.org/2013/11/08/the-deep-web*
4. *Lederman, Abe and Lederman, Sol. "Understanding Deep Web Technologies." New Idea Engineering. Jun. 2004. (Dec. 6, 2013) http://deepwebtech.com/PDFs/Understanding%20Deep%20Web%20Technologies.pdf*

5. *Swift, Tim. "What is the 'Deep Web'? And Other Questions About the Shadowy Virtual World of Silk Road." Baltimore Sun. Oct. 3, 2013. (Dec. 6, 2013) http://articles.baltimoresun.com/2013-10-03/business/bal-silk-road-deep-web-explainer-20131003_1_satoshi-nakamoto-bitcoin-silk-road*
6. *University of California, Berkeley. "Invisible or Deep Web: What It Is, How to Find It, and Its Inherent Ambiguity." (Dec. 6, 2013) http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html*
7. *http://en.wikipedia.org/wiki/Tor_(anonymity_network)*
8. *https://cryptogasm.com/2012/08/deep-web-dark-web/*