Dark Web:
A Boon or a Bane

Punam Bedi
https://orcid.org/0000-0002-6007-7961
University of Delhi, India

Neha Gupta
University of Delhi, India

Vinita Jindal
https://orcid.org/0000-0002-0481-4840
Keshav Mahavidyalaya, University of Delhi, India

INTRODUCTION

Ever since its conceptualisation, the Internet has revolutionised the way people live. From faster communication to convenient information sharing, the human race has been abundantly blessed by the applications of the Internet. Most people today, mainly use the Internet for accessing various websites or web pages containing different forms of information – like text, image, audio, video etc. Such widespread use of the Internet for data retrieval purpose, has got people thinking about the difference between the Internet and the World Wide Web (or simply the Web). In essence, the Internet comprises of networks across the world, interconnected together to form one massive global network. These networks are connected via cables or satellite links and follow a predefined set of rules for communication, called protocols. The Internet uses different protocols to deliver a variety of services to its users – like data dissemination, email, file sharing, online messaging etc. Out of all such services the World Wide Web only provides data dissemination or information sharing capability to the Internet. It allows web pages to display content present on world's different networks by using Hyper Text Transfer Protocol (HTTP). Services like email and file sharing, which are not a part of the World Wide Web, are governed by Simple Mail Transfer Protocol (SMTP), Post Office Protocol (POP) and File Transfer Protocol (FTP) respectively. Since users mostly access the Web portion of the Internet to share and/or retrieve information, the two terms – the Web and the Internet - are used interchangeably by common people. But these two terms are not synonymous. The Internet is the underlying (hardware) infrastructure over which the World Wide Web (software) hosts digital content in the form of websites or simply web pages.

SURFACE WEB

The Web provides its users with billions of web pages that can be easily accessed via standard web browsers and search engines. This part of the Web which is openly available to everyone is known as the Surface Web (Santos, 2015). Users access online resources present on Surface Web by either typing

DOI: 10.4018/978-1-5225-9715-5.ch010

a Uniform Resource Locator (URL) in the web browser or submitting a string of keyword(s) to query the search engine. The latter option provides them with an ordered list of search results that are most relevant to their search query. Both these methods allow quick and easy access to information on the Surface Web - also known as Visible Web or Clearnet. With huge amounts of information lying just a few clicks away, the general audience tends to think that all online content hosted on the World Wide Web is part of the Surface Web and can be accessed by them through conventional search engines (like Google and Bing) and web browsers. But this is far from the truth. The Surface Web contains less than 20 percent of the total information present on the Web (Santos, 2015) (Sui, Caverlee, and Rudesill, 2015). This is because the Visible Web is solely formed by the contents that search engines are able to reach on the Web. Any web resource that is beyond the reach of search engines, is not a part of the Visible Web. Thus the volume of content present on the Surface Web is limited by the techniques that search engines follow to extract information from the World Wide Web.

Search engines make use of application programs that scan the World Wide Web to create an index of all "reachable" web resources. These programs are known as crawlers/spiders/harvesters. Crawlers navigate the Web to gather documents and files present online in the form of web pages. Usually, web pages are linked to each other via incoming and outgoing hyperlinks. These hyperlinks enable search engines' spiders to reach different web pages and extract information from them. This process of gathering information while moving from one web page to another through hyperlinks, is known as crawling. Beginning with an initial set of seed URLs, a crawler scans every web page linked to these URLs via outgoing hyperlinks (Santos, 2015). All online content that is reachable via incoming and/or outgoing hyperlinks gets crawled. The crawled pages are then indexed to allow easy retrieval for later purposes. Finally, when the user submits a search query, the search engines display these indexed web pages in an ordered manner.

Though search engines have made a lot of information available, but a very large portion of the Web still remains inaccessible to users. This is because web spiders can only enter and leave web pages via hyperlinks. Web pages which do not have incoming hyperlinks cannot be reached by crawlers. Such pages remain un-crawled and their content remains hidden from the eyes of search engines' spiders. This is why the number of web pages on the Surface Web is a subset of web pages available on the World Wide Web. The Web is an ocean of knowledge and crawlers dive into it to gather as much content as possible. Contents gathered by crawlers form the Surface Web while the remaining contents form a separate layer of the Web, called the Deep Web.

DEEP WEB

The World Wide Web is a collection of enormous data both accessible directly and not accessible directly through crawlers. The accessible data, which is indexed by standard search engines, forms the visible tip of the iceberg whose major portion lies unseen in the deeper layers of the Web. This "unseen" portion of the Web which cannot be indexed by traditional search engines is referred to as the Deep Web. It consists of all those contents that cannot be crawled by search engines' spiders and hence does not appear in search results. Though Deep Web content remains hidden from regular searches, it is not usually illegal or dangerous, it is just made to be unsearchable. Most of this information can be displayed as a result of a targeted search over the Web. Deep Web, popularly known as the Hidden Web or Invisible Web,