Mining the Assets of Invisible Web

Ms. Geet Bawa
Assistant Professor,
PG Department of Computer Science
Khalsa College for Women (Autonomous),
Amritsar 143001
Punjab

Abstract:

The World Wide Web is increasingly integral to social, cultural, political, educational, academic, and commercial spheres. It hosts a vast array of information and applications relevant to society. Many users rely on search engines for information retrieval but often hesitate before making final decisions due to the limited and imprecise information available. Millions of high-quality resources on the web remain unseen by general-purpose search engines. This may result from using irrelevant keywords or selecting inappropriate search engines. Frequently, search engines fail to pinpoint exactly what users seek. Beyond these issues, a significant reason for inefficient search results is the technical limitations of search engines in accessing and retrieving certain web content. Some information is hidden from even the most efficient search engines, known as the "Invisible Web." The Invisible Web includes resources not indexed by standard search engines, but this does not render them unimportant. Information not accessed by search engines is as valuable as the accessible content. The Invisible Web is a noteworthy phenomenon. This paper explores the concept of the Invisible Web and examines why search engines cannot access all web content. It also discusses the various resources within the Invisible Web and provides a list of search engines capable of mining and harvesting this hidden content.

Keywords: Search Engines; Invisible Web; Surface Web; Internet Portals.

Introduction

Search engines have become an indispensable part of our daily lives, seamlessly integrating into our routines. We rely on these tools to navigate the vast expanse of the internet. A search engine is a software application that searches its database of websites based on the keywords we input and returns a comprehensive list of web addresses containing the relevant information. There are numerous search engines and internet directories available, such as InfoSeek, Google, Yahoo!, Excite, HotBot, AltaVista, Lycos, and LookSmart, to name a few. Many of these major search engines are evolving into internet portals. Given that search engines are the primary means of locating information on the web, it is crucial to understand how to use them effectively and efficiently.

Related Work

Most of the information available on the internet often cannot be retrieved by search engines. This subset of the web, known as the "Invisible Web," includes files, text pages, and websites that search engines do not index due to technical limitations. In essence, the Invisible Web encompasses all the information on the World Wide Web that general-purpose search engines cannot find (Devine and Egger-Sider, 2001). The Invisible Web poses significant challenges for the information world. The term "Invisible" implies obscurity and marginalization, leading some to prefer terms like "Deep Web," "Dark Matter," or "Hidden Web" to describe such content. A prominent analogy illustrating the relationship between the Invisible Web and the Visible Web is that of a fishing trawler with its net cast in the middle of the ocean (Bergman, 2001). The ocean represents the entirety of information available on the World Wide Web. The depth reached by the net symbolizes the content captured by general-purpose search engines, known as the Visible Web or "surface web." The ocean beyond the net's reach represents the Invisible Web. Thus, the Visible Web and the Invisible Web are parts of the same vast information world as highlighted in Figure 1.

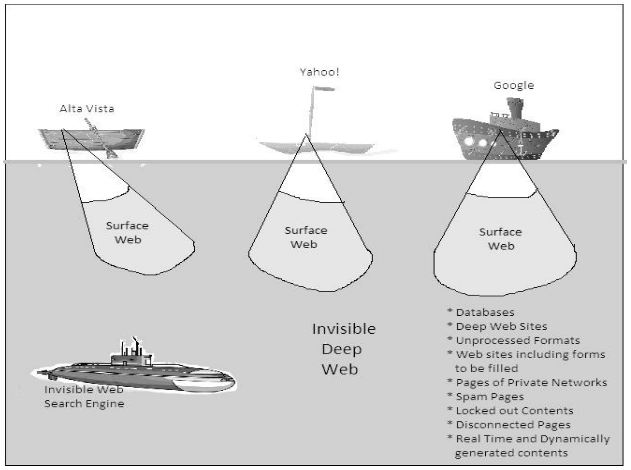

Figure 1: Surface Web and Invisible Web

Need and Importance of the Study

The Invisible Web represents the largest sector of online information resources on the internet. To locate such information, one must know exactly where to search. Search engines like Google and Yahoo! provide access to only a small fraction of web information. Studies have shown that the Invisible Web is approximately 500 times larger than the visible web accessed by general-purpose search engines like Google (Bergman, 2001). The immense size of the Invisible Web compared to the visible web presents a significant challenge. This discrepancy arises from the limitations of search engine technologies, which, despite offering vast information, cannot access all web content. Additionally, the high costs of operating search engines, including fetching web resources and maintaining up-to-date indices, contribute to this issue. Given that a substantial amount of crucial information resides on the Invisible Web, one might ask, "Why are some web contents invisible?" To answer this, it's essential to understand the properties and boundaries of the Invisible Web. The Invisible Web includes all web content that cannot be accessed using general-purpose search engines, highlighting a direct relationship between the Invisible Web and these search engines. Each search engine effectively creates its own Invisible Web, consisting of information it does not index (Sullivan, 2008). Consequently, the size of the Invisible Web varies from one search engine to another. Currently, no existing search engine can access the entire world of web information.

The Problem Statement: Assets Present in the Invisible Web

Search engines exclude a vast collection of information due to various practical and technical considerations, such as formats, size, and ease of indexing. Here, we will examine the different types of resources found in the Invisible Web.

Databases

Databases are dynamically generated collections of information, which makes it challenging for search engine spiders to access and retrieve data from them. While search engines can identify the homepage of a database, they cannot delve into its contents. When a query is made in a database, it is processed dynamically by the database program, and the results are displayed. If these results are no longer needed, they are disassembled. There are no pre-computed answers that can be displayed directly and quickly, making the outputs dynamic rather than fixed, and thus, unrecognizable for future reference. To achieve the same results again, the user must reconstruct the query. Databases rely on their own search programs to generate output results, and these search procedures are specific to each database. The majority of the Invisible Web is composed of these databases, as mentioned in Table 1. Databases offer a flexible and easily maintainable development environment for their creators. However, each database is unique in terms of its data structure design, search tools, and capabilities, making it a significant challenge for any search engine to work effectively with databases.

Table 1: List of top 10 largest deep web sites among 60 sites (arranged in descending order of their size in GBs)

S. No.	Name of the site	URL	Type of site	Website size (GBs)
1.	National Climatic Data Center (NOAA)	http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html	Public	366,000
2.	NASA EOSDIS	http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html	Public	219,600
3.	National Oceanographic (combined with Geophysical) Data Center (NOAA)	http://www.nodc.noaa.gov , http://www.ngdc.noaa.gov/	Paid/Public	32,940
4.	Alexa	http://www.alexa.com/	Public	15,860
5.	Right-to-Know Network (RTK Net)	http://www.rtk.net/	Public	14,640
6.	MP3.com	http://www.mp3.com/	Public	4,300
7.	Terraserver	http://terraserver.microsoft.com/	Paid/Public	4,270
8.	HEASARC (High Energy Astrophysics Science Archive Research Center)	http://heasarc.gsfc.nasa.gov/W3Browse/	Public	2,562
9.	US PTO - Trademarks +	http://www.uspto.gov/tmdb/ ,	Public	2,440

	Patents	http://www.uspto.gov/patft/		
10.	Informedia (Carnegie Mellon Univ.)	http://www.informedia.cs.cmu.edu/	Public	1,830

Deep Web Websites

The World Wide Web hosts many websites that are rich in resources and content. One of the technical limitations of any general-purpose search engine is its depth of crawl. Search engines set limits on how much content and how many pages they index from a site, resulting in the exclusion of extensive and deep websites that may contain valuable information. As websites grow, the amount of excluded content also increases. Examples of such rich, complex sites include government websites like the Library of Congress (www.loc.gov) and the Census Bureau (www.census.gov). According to a report by BrightPlanet, a company specializing in deep Invisible Web research for the business world, the 60 largest deep web sites contain data equivalent to 40 times the information found on the visible web. BrightPlanet produced a list of these 60 largest deep web sites, highlighting the vast resources available beyond the reach of standard search engines (www.brightplanet.com/inforcenter/largest_deepweb_sites.asp).

Unprocessed Formats

Search engines are typically designed to process a limited number of file formats, often excluding many other formats. When a new format becomes available on the Internet, search engines must either adjust their spiders' programming or develop new search procedures to index pages with these new format contents. More often than not, search engines omit such content, contributing to the growth of the Invisible Web. In other words, most current search engines are designed to index and process text. When they encounter non-textual files or objects, their performance declines significantly. The World Wide Web (WWW) may be the largest repository of digital images globally. The number of images available on the Internet is rapidly increasing and will continue to grow. Users need efficient tools to browse and search for these images. Current image search engines, such as TinEye, can partially fulfill this need. However, these image processing search engines often fail to find all similar images on the web, as they rely on visual characteristics like color, texture, and shape for comparison. Some search engines, like

AltaVista and HotBot, are designed to perform non-textual searches, including images, audio, and video files, but they still face significant limitations. Thus, pages consisting mainly of audio, video, images, or compressed files (.zip, .tar, etc.) with little or no text form a substantial part of the Invisible Web.

Websites which include Forms to be filled

Certain websites, aside from database sites, feature forms that users fill out to generate dynamic information. These customized contents pose challenges for search engine spiders similar to those encountered with databases. For example, job search sites require information such as the user's location and interests. Once the user fills out the form with the necessary details, the site generates a response tailored to the user's query. This response is dynamically created for that specific user and disappears once the user finishes with it. This transient information, generated by such sites, also constitutes a significant portion of the Invisible Web due to its ephemeral and real-time nature.

Other Resources

Sometimes the owners of web sites do not want their confidential information to be visible on search engines. These include the pages that belong to private networks of organization. Other reasons include the strictness done by search engine to deal with spam pages which is unfortunately the reason for excluding the legitimate information. These all lead to problem of Invisible Web. Moreover, there are web contents that the search engines have decided to exclude such as all the first-rate content sources which are effectively locked out web contents. These include library databases which need a password to access, as mentioned in Table 2. In addition to this, search engine uses a program known as crawler to retrieve web pages stored on servers all over the globe. These crawlers rely on the links present on the pages to access other pages. So the limitation is that if there is a web page which has no link pointing to it from any other page on the web, search engine crawler cannot find it. These disconnected and unreached pages are the major part of Invisible Web [6] [7] [8].

Table 2: Resources of Invisible Web along with the reasons of their invisibilities

Resources of Invisible Web	Reasons for Invisibility

Databases	Databases rely on their own search procedures to request output results.
Deep Web Sites	Search engines impose limits on the amount of content and number of pages they index from a site.
Unprocessed Formats	Pages containing images, audio, video, PDFs, Flash, Shockwave, or compressed files may have limited textual data, making indexing difficult.
Websites including forms to be filled	Short-lived nature of content.
Pages of private networks	Confidentiality of information restricts indexing.
Spam pages	Stringent restrictions enforced by search engines.
Locked out web contents	Password requirements to access contents prevent indexing.
Disconnected pages	Lack of inbound links from other pages.
Real time contents	Rapidly changing and ephemeral nature of content.
Dynamically generated contents	Customized information that disappears after a period of time.

Solution: Use of Search Engines that could Mine Invisible Web

Unlike general-purpose search engines such as InfoSeek, Google, Yahoo!, Excite, HotBot, AltaVista, Lycos, and LookSmart, there exist many search engines that function as specialized deep-diving vessels for the Invisible Web, as presented in Table 3. These Invisible Web search engines index specifically targeted information. Researchers and educators should prioritize these search engines over general-purpose ones for indexing the vast contents and information within the realm of the Invisible Web.

Table 3: List of few Invisible Web Search Engines

Name of Search Engine	URL
Complete Planet	http://www.brightplanet.com/completeplanet/
DeeperWeb	http://deeperweb.com/
DeepPeep	http://org.deeppeep.qirina.com/

DeepWebTech	http://www.deepwebtech.com/
Dogpile	http://www.dogpile.com/
Factiva	https://global.factiva.com/factivalogin/login.asp?productname=global
FindArticles	http://www.search.com/search
FindSmarter	http://findsmarter.com.hypestat.com/
Forrester Research	https://www.forrester.com/home/
Harvard	http://adswww.harvard.edu/
IncyWincy	http://www.incywincy.com/
Infomine	http://library.ucr.edu/view/infomine
Infoplease	http://www.infoplease.com/
Library of Congress	https://catalog.loc.gov/
National Security Achieve	http://nsarchive.gwu.edu/search.html
Navagent	http://www.navagent.com/
Quintura	http://quinturakids.com/
Surfwax	http://lookahead.surfwax.com/
TechXtra	http://www.techxtra.ac.uk/
The WWW Virtual Library	http://vlib.org/
TouchGraph	http://www.touchgraph.com/navigator
US Geologic Survey	http://search.usgs.gov/
Xrefer	http://www.xrefer.com/
Yippy	http://yippy.com/
Zuula	http://www.zuula.com/

Conclusion

The Invisible Web constitutes a vast repository of valuable content crucial for research across various domains. Such resources demand significant attention and should not be overlooked. Serious information seekers recognize the significance and value of the contents within the Invisible Web. Moreover, the Invisible Web is expanding at a rapid pace comparable to that of the visible web. By comprehending the potential, advocating for, and educating others about the

Invisible Web, information professionals can harness its resources and contents effectively. Utilizing search engines specifically developed for harvesting the Invisible Web presents a superior solution. Furthermore, given that current general-purpose search engines are continuously updating and enhancing their services, it is plausible to anticipate that what is invisible today may become visible tomorrow.

References

Bales, R. A., & Stone, K. V. (2020). The Invisible Web at Work. *Berkeley Journal of Employment and Labor Law*, *41*(1), 1-61.

Alyami, H. Y., & Assiri, E. A. (2018). Invisible Web and Academic Research: A Partnership for Quality. *International Education Studies*, *11*(4), 84-91.

He, S., He, Y., & Li, M. (2019, March). Classification of illegal activities on the dark web. In *Proceedings of the 2nd International Conference on Information Science and Systems* (pp. 73-78).

Nazah, S., Huda, S., Abawajy, J., & Hassan, M. M. (2020). Evolution of dark web threat analysis and detection: A systematic approach. *Ieee Access*, *8*, 171796-171819.

Bergman, J., & Popov, O. B. (2023). Exploring dark web crawlers: a systematic literature review of dark web crawlers and their implementation. *IEEE Access*.

Ofusori, L., & Hendradi, R. (2023). Understanding the Impact of the Dark Web on Society: A Systematic Literature Review. *International Journal of Information Science and Management (IJISM)*, *21*(4), 1-21.

Basheer, R., & Alkhatib, B. (2021). Threats from the dark: a review over dark web investigation research for cyber threat intelligence. *Journal of Computer Networks and Communications*, *2021*, 1-21.

Alaca, F., Abdou, A., & van Oorschot, P. C. (2019). Comparative analysis and framework evaluating mimicry-resistant and invisible web authentication schemes. *IEEE Transactions on Dependable and Secure Computing*, *18*(2), 534-549.

Odey John, A., & Okoro Anthony, T. ACCESSING THE INVISIBLE WEB: ISSUES AND CONCERNS.

Marjan, M. (2019). Illegal Activities in the Invisible Web. *Crim. Just. Issues*, 87.

Zink, R., Sarapura, S., Potter, C., Fontecha, M., & Cupolo, N. (2024). Within the invisible web: Gender-based violence in agricultural streams of Canada's Temporary Foreign Worker Program. *Rural Review: Ontario Rural Planning, Development, and Policy*, *8*(1).

Sohrabi, M. C. Scientists' use of visible and invisible Web: an analysis based on Max Weber's ideal type.

Abdellatif, T. M., Said, R. A., & Ghazal, T. M. (2022, October). Understanding Dark Web: A Systematic Literature Review. In *2022 International Conference on Cyber Resilience (ICCR)* (pp. 1-10). IEEE.

Susuri, A. (2019). Dark web and its impact in online anonymity and privacy: A critical analysis and review. *Journal of Computer and Communications*, *7*(3), 30-43.

Kaur, S., & Randhawa, S. (2020). Dark web: A web of crimes. *Wireless Personal Communications*, *112*, 2131-2158.

Faizan, M., & Khan, R. A. (2019). Exploring and analyzing the dark Web: A new alchemy. *First Monday*.