IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-3, Issue-11 E-ISSN: 2347-2693

Harvesting the Resources of Invisible Web

Hardeep Singh[1*] and Geet Bawa[2]

[1]Post Graduate Department of Computer Science, BBK DAV College for Women, Amritsar, India
[2]Post Graduate Department of Computer Science, Khalsa College for Women, Amritsar, India

www.ijcseonline.org

Received: Oct/22/2015 Revised: Nov /05/2015 Accepted: Nov/20/2015 Published: Nov/30/2015

***Abstract*—**The World Wide Web is constantly becoming an important part of social, cultural, political, educational, academic, and commercial life. Web contains a wide range of information and applications in areas that are of societal interest. A great number of World Wide Web users use search engines for information retrieval, but still hesitate before making a final decision, often because only rough and limited information about the products is made available. There are millions of high quality resources available on web that the general-purpose search engines can't see. One of the supportive reasons for this could be use of irrelevant keyword(s) or choice of a wrong search engines for executing a particular request of the searcher. Many times search engine cannot find out what we exactly wanted from it. The major reason why sometimes we do not succeed to acquire efficient results, other than these reasons, is the technical inability of search engines to access and retrieve some of the contents present on the web. That is, some of the information is hidden from the eyes of even efficient search engines. Such information which remains inaccessible from web search engines is termed as "Invisible Web". Invisible Web contains resources that are not indexed by general-purpose search engines, but this does not indicate that these resources are absolute leftovers and unimportant. The information that is not accessed by a search engine is as much significant as that which is accessed. Invisible web is a phenomenon to be reckoned with. This paper provides a view of Invisible Web and also delves into the reasons why search engines can't see all of the web contents. Various resources present in invisible web are also discussed. Paper also provides a list of search engines that could mine and harvest Invisible Web.

***Keywords*—**Search Engines; Invisible Web; Surface Web; Internet Portals.

I. Introduction

Search engines have become a part of our daily lives. We cannot imagine our daily routine without these pathfinders. A search engine is software installed on the internet that searches its database of websites based on the keyword we enter and shows a huge list of web addresses which contain the information we are searching for. There are large number of search engines and internet directories available including InfoSeek, Google, Yahoo!, Excite, HotBot, AltaVista, Lycos, LookSmart etc to name a few. Many of the major search engines are becoming known as Internet Portals. Since search engines are the major tools in locating desired information on the web, it is important to know how to use them effectively and efficiently.

II. RELATED WORK

The information present in the internet cannot be retrieved by search engines most of the times. Such information that is available on the web but is not easily located by the general-purpose search engines is collectively known as "Invisible Web". Invisible web actually include files, text pages or sites which are available on World Wide Web but the search engines do not index them due to some technical limitations. Simply put, Invisible Web is used to describe all the information available on World Wide Web that cannot be found using general-purpose search engines (Devine and Egger-Sider, 2001). Invisible web presents many problems for the information world. The term "Invisible" implies obscurity and marginalization, due to which it is sometimes preferred to use the term "Deep Web" or "Dark Matter" or even "Hidden Web" to refer to such contents [1]. One of the prominent images that could reveal the relationship between invisible and visible web is one of the fishing trawler with its net out in the middle of the ocean (Bergman, 2001). The ocean can be imagined as a collection of whole information that is available on World Wide Web. The depth of the ocean accessible by the net represents the contents that are captured by the general-purpose search engine, which is the visible web, also known as "surface web". The ocean beyond the net represents the Invisible Web. Hence "Visible Web" and "Invisible Web" are actually the parts of same world of information [2].

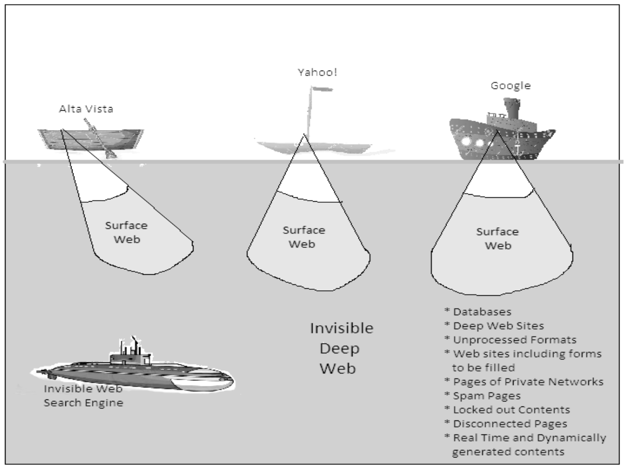

***Figure 1**: Surface Web and Invisible Web*

III. NEED AND IMPORTANCE OF STUDY

The Invisible Web represents the largest sector of online information resources on the internet. In order to locate such information, there is a need to know exactly where that information is to be searched. Search engines like Google and Yahoo! give us view of only small fraction of web accessible information. Studies have shown that Invisible Web is about 500 times the size of visible web accessed by any general-purpose search engine such as Google (Bergman, 2001). The size of Invisible Web dwarfs the size of visible web which is the major challenge to be considered. The reason is that search engine technologies are limited in their capabilities, despite providing vast information to their searchers. Moreover, the costs involved in operating search engines is also an eminent factor as it is expensive for any search engine to fetch web resources and retain up-to-date web indices. Since large amount of important and major information resides on invisible net, so question that arise is "Why some web contents are invisible?" To answer this, need to get a clear picture of properties and boundaries of invisible net. Invisible Web comprises of all the contents available on web that cannot be accessed by using general-purpose search engines, so there is a direct relationship between Invisible Web and general-purpose search engines. We can propose that every search engine creates its own Invisible Web which is basically the collection of information that is not indexed by that search engine [3]. Therefore, not only does each search engine creates its own Invisible Web of excluded items, but also the size of that web varies from one search engine to another (Sullivan, 2008). Currently, none of the existing search engines are able to access whole world of web information.

IV. PROBLEM STATEMENT: RESOURCES PRESENT IN THE INVISIBLE WEB

Search engines exclude huge collection of information due to some practical and technical considerations including formats, size, ease in indexing etc. Here we will examine different types of resources found in Invisible Web.

A. Databases

Databases are dynamically generated collection of information. Due to this property, it becomes problematic for search engines spiders to enter into and retrieve data from them. Search engines can definitely identify database homepage but cannot seek database contents. Whenever a query is made in a database, it is dynamically processed by the database program and the result outputs are displayed. If those results are no more needed then they are disassembled. There are no pre-computed answers that could be displayed directly and quickly. The outputs are dynamic entity; rather than fixed entity, hence cannot be recognized again. To achieve same results in future, user needs to restructure the query. Databases depend on their own search programs to request for output results and these search procedures are database dependent. Majority of the Invisible Web is created by databases. Databases provide a flexible and easily maintainable development environment for their creators. Since every database is unique in terms of design of its data structure, its search tools and capabilities, so working with databases is a real challenge for any search engine.

B. Deep Web Websites

World Wide Web contains many web sites that are very deep and rich in resources and contents. One of the technical limitations in any general purpose search engine is its depth of crawl. Search engines put a limit on how much content and how many pages they index from a site. This limits results in exclusion of very rich, extensive and very deep web sites, which may contain important information. Moreover, as the web sites grow the excluded contents also grow. Examples of such rich complex sites include government web sites such as Library of Congress (www.loc.gov) and the Census Bureau (www.census.gov). According to a report generated by BrightPlanet, a company that provides a research of deep Invisible Web for the business world, a list of 60 largest deep web sites was produced.BrightPlanet(www.brightplanet.com/inforcenter/largest_deepweb_sites.asp) also explored that resources of these 60 web sites are equal to 40 times the data found on visible web [4].

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)	http://www.nodc.noaa.gov/ , http://www.ngdc.noaa.gov/	Paid/Public	32,940

	Data Center (NOAA)			
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 + Patents	http://www.uspto.gov/tmdb/ , http://www.uspto.gov/patft/	Public	2,440
10.	Informedia (Carnegie Mellon Univ.)	http://www.informedia.cs.cmu.edu/	Public	1,830

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

C. Unprocessed Formats

Search engines are usually designed to process limited number of file formats. Many of the formats are excluded by the search engine crawlers. Moreover, when a new format is available on the Internet, search engines need to either adjust their spider's programming or develop new special search procedures to index pages including new format contents. Mostly search engines usually omit such contents resulting in enrichment of Invisible Web. In other words, most of the current search engines are designed to index and process text, so when they encounter a file or an object that is non-textual in nature they discontinue performing well. The Word Wide Web (WWW) may be the largest repository of digital images in the world. The number of images available on the Internet is increasing rapidly and will continue to grow in the future. There are so many images available on the Internet that users do need efficient tools to browse and search for those images. The current image search engines, for example TinEye search engine, can partially fulfill this need [5]. The current image processing search engines fails to find all the images from web which are similar to the image given as a query by comparing visual characteristics, such as color, texture and shape of the input image. Some of the SE like AltaVista and HotBot are designed to do certain non-textual search such as images, audio and video files but there are still certain limitations. So the pages that consist mainly of audio, video, images or compressed files (.zip, .tar etc) with little or no text are a major part of Invisible Web.

D. Websites which include forms to be filled

Some web sites, other than database sites, which include forms to be filled by the user, generate dynamic information. Such customized contents present problems for search engine spiders similar to those of databases. For instance, job searching sites must know the location and interests of the job finder. Once the form inquiring needed information is filled by the user, the site generates response to the user's query. This response is created dynamically for that specific user and vanishes when the user is finished with it. This short-lived information which is created by such sites also forms a major part of Invisible Web due to its ephemeral and real time nature.

E. 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. 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].

Resources of Invisible Web	**Reasons for Invisibility**
Databases	Dependence of databases on their own search procedures to request for output results
Deep Web Sites	Limit imposed by search engines on how much content and how many pages they index from a site
Unprocessed Formats	Pages containing images, audio, video, PDF, Flash Shockwave or compressed files due to limited textual data in them
Websites including forms to be filled	Due to their short-lived nature
Pages of private networks	Due to confidentiality of such information
Spam pages	Due to the strictness done by search engines
Locked out web contents	Due to the requirement of passwords to open such contents
Disconnected pages	As these pages does not links pointing towards them on other pages
Real time contents	Due to its rapidly changing ephemeral nature
Dynamically generated contents	Customized information which vanishes after a period of time

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

V. OBJECTIVE: USING SEARCH ENGINES THAT COULD MINE INVISIBLE WEB

Unlike general-purpose search engines like InfoSeek, Google, Yahoo!, Excite, HotBot, AltaVista, Lycos, LookSmart etc to name a few, there are many search engines that could be used as a deep diving vessel for the Invisible Web. These are the Invisible Web search engines with specifically indexed information. Researcher and educationalist should prefer these search engines over general-purpose search engines to index the contents and information present in the world of Invisible Web.

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/

***Table 3**: List of few Invisible Web Search Engines*

CONCLUSION

Invisible Web is a large source of valuable web that is important for research on many levels. Such resources should not be overlooked and needs major consideration. Serious information seekers cannot elude the significance and value of contents present in Invisible Web. Invisible Web also grows at the faster rate as the visible web. By understanding the prospective, promulgation and teaching Invisible Web, the informational professional can brings its resources and contents into use. Using search engines developed for harvesting Invisible Web is the better solution. As the current general-purpose SE are constantly updating and improving their services, so we can conclude that what is invisible today may be visible tomorrow.

REFERENCES

[1] Jacsó, P. (2005), "Google Scholar: the pros and cons", *Online Information Review*, Vol. 29, No. 2, pp. 208-214.

[2] CompletePlanet. (2004). “Largest deep web sites”. BrightPlanet. Available: http://aip. completeplanet.com/aip-engines/help/largest_engines.jsp

[3] Devine, Jane, and Francine Egger-Sider. 2001. *Beyond Google: The Invisible Web*. Available: www.lagcc.cuny.edu/library/invisibleweb/definition.htm

[4] Bergman, Michael K. (2001). “The deep Web: Surfacing hidden value.” White paper. BrightPlanet. Available: www.brightplanet.com/images/stories/pdf/deepwebwhite paper. pdf

[5] Sullivan, Danny. (2008). “Google now fills out forms and crawls results.” Search Engine Land. Available: http://searchengineland.com/080411-140000.php

[6] Williams, M.E. (2005), "The state of databases today: 2005", in *Gale Directory of Databases*, Vol. 2, pp. XV-XXV, Gale Group, Detroit, MI.

[7] Ru, Y. and Horowitz, E. (2005), "Indexing the invisible web: a survey", *Online Information Review*, Vol. 29, No. 3, pp. 249-265.

[8] Calishain, Tara. 2005. “Has Google dropped their 101K cache limit?” ResearchBuzz! Available: www.researchbuzz.org/2005/01/has_google_dropped_their_101k.shtml