First International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, 22-25 October, 2009

Invisible Web Search: Case Study Deep Web Search Tools

IOAN POP

Abstract

Some users have the sensation that they can found anything on the World Wide Web by using common search engines. This approach is not correct, as powerful as these search engines are, they do not index everything on the World Wide Web. This is the massive content that is publicly available, but hidden from regular search engines. The spiders or crawlers do not have the capability to index a big part of the web documents. In order to find the information from the web document repository it is necessary the special techniques and resources. In this paper we propose some approaches to mining the invisible web through: some strategies to access and mine the invisible web; many web search tools; techniques and resources to search the deep web; webometrics for the deep web.

1 Introduction

The term "Invisible Web" (or "Deep Web" or "Cloaked Web") is recently accepted by the people that research the web. Invisible web mainly addresses to the growth repository of the documents that you can't access with search engines and directories. To be specific: the Invisible Web is comprised of hundreds billion web pages that are not stored as static web pages. Instead, the Invisible Web is made of on-demand database content pages which exist only as reports of changing data. Today, robot crawlers are not advanced enough to read these private databases. Only a human reader can see these "invisible pages" by directly visiting these sites and making direct database requests [7].

To dig deeper into the Web, a new breed of search engine has cropped up that takes a different approach to Web page retrieval. Instead of broadly scanning the Web by indexing pages from any links they can find, these search engines are devoted to drilling further into specialty areas i.e.: medical sites, legal documents, even Web pages dedicated to jokes and parody. A few search engines have tried to take that step, with mixed results. But again, many Web users do not know that the narrow searching tools exist. So reference librarians and library Web sites are now directing their patrons to those areas on the Web.

BrightPlanet is the leader in harvesting high quality content from inaccessible Deep Web and Surface Web sources. With over 10 years of Deep Web extraction expertise, the company has developed a heuristic, rule-based expert system for communicating with Deep Web sources that does not require one-off scripts to be built by hand, which are often prone to failure. The fully automatic configuration system configures about 80% of known Deep Web sources without any user intervention. Another 15% of Deep Web sources can be configured with only minor user intervention, requiring only about 5 seconds per source. All remaining Deep Web sources can be configured using an extensive scripting language that also supports password protected and JavaScript sites [7].

A recent study, BrightPlanet's white paper indicates that the invisible web is 400-550 times larger than the traditional (surface or open) web. They estimate that there are more than 7,500 terabytes of information contained in the invisible web, compared to the 19 terabytes of the surface web. Most of these sites are accessible to the public, free of charge [2].
Major components of the invisible web include: non-HTML files (i.e.: PDF files, Flash files, etc.), sites requiring registration of login, archives (magazines, newspapers, etc), interactive tools (calculators, translators, etc.).

2 Strategies to access and mine the invisible web

The researchers often need more than Google and Wikipedia to get the job done. To find what you're looking for, it may be necessary to tap into the invisible web, the sites that don't get indexed by broad search engines. The following resources were designed to help you do just that, offering specialized search engines, directories, and more places to find the complex and obscure.

2.1 A stratified view of the web

An illustration of the Web "*Content Layers*" is presented in the figure 1.

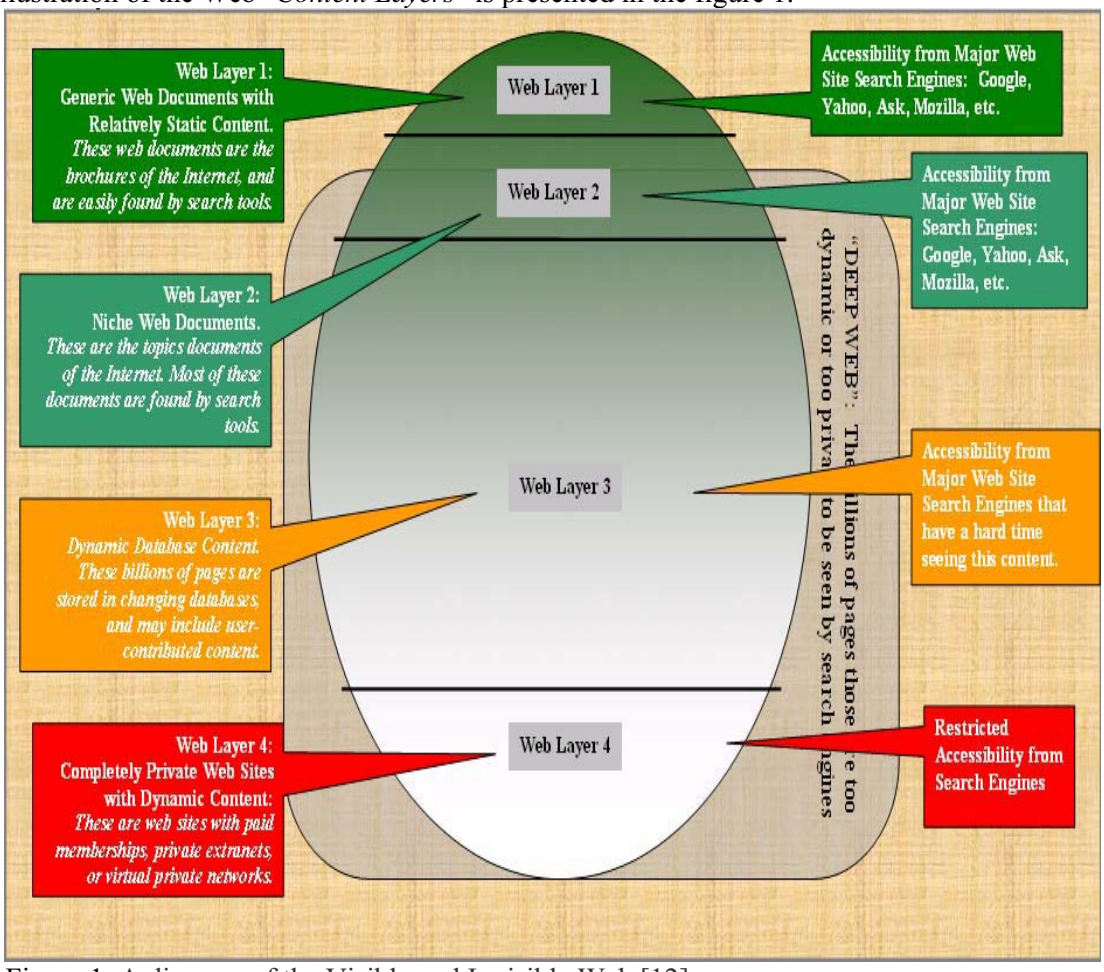

Figure 1. A diagram of the Visible and Invisible Web [12].

There are two pole tips of the Web: **a**. the surface web (layer 1 and layer 2) that is visible by the commonly search engines; **b**. the invisible web (layer 3 and layer 4) that is not visible by the commonly search techniques. The content of the deep web can only be found by sending a direct query to a database. Traditional search engines are not able to detect the contents of these databases as their crawlers are collecting information for their catalogs. However, more and more traditional search engines are adding deep web searching capabilities to their sites.

2.2 Deep Web Search Strategies

Currently there are two ways the deep web can be searched: on the one hand we want instant access to information throught a specialized deep web search tools, and on the other hand throught a specialty search engine passed on to you.
First, let's fin some deep web search engines. Some are easy to find in the surface web by going to a search engine and searching "Deep Web search engines". From time to time you will pick up a good deep web site by "word of month", at a conference or in a deep web class such as this one. These sites may or may not show up in a normal deep web search because of their specialization or newness, or they may be password protected.
For example, the steeps that can be followed in the try to mine the deep web are the following: **a**. being aware that the deep web exists; **b**. using a general search engine for broad topic searching; **c**. using a searchable database for focused searches; **d**. register on special sites and use their archives; **e**. call the reference desk at a local college if you need a proprietary web site; **f**. many college libraries subscribe to these services and provide free on-site searching (and a friendly trained librarian to help you); **g**. check the web site of your local public library; **h**. many libraries offer free remote online access to commercial and research databases for anyone with a library card.

3 Techniques for Search the Deep Web

The web sites that usual search engines can't find is the following types of: sites with dynamic scripting, private web pages, sites that require a registration, temporary web pages, blocked sites by local webmasters, sites blocked by search engine policy, sites with special formats, searchable databases [18].

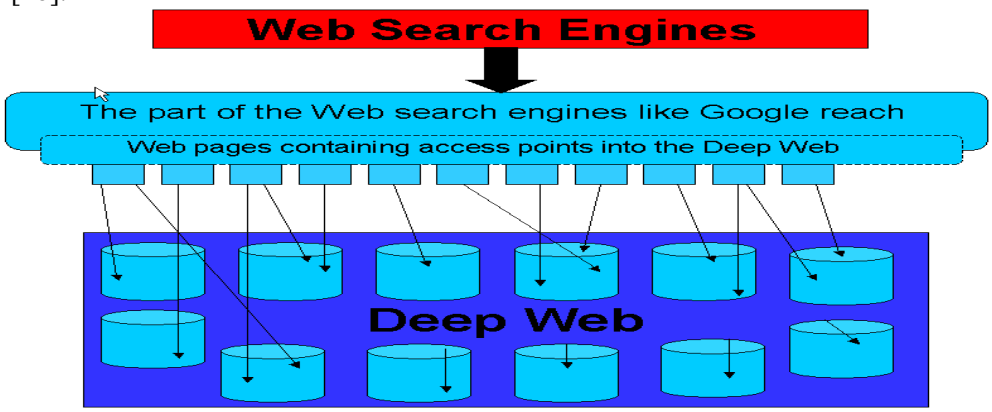

Figure 2. A diagram of the Deep Web Search Engines[14].

3.1 Searching the Invisible Web

In order to search the invisible web the search engines vendor now make tools available. Rather than retrieving web pages or documents, invisible web search engines direct the user to an appropriate searchable database. Some even generates a search form for your convenience. Searching the invisible web is processes in two-steps. According [19] these two steps are:
First, locate the appropriate database to search. Browsing is the easiest method. Invisible web search engines are accompanied by a categorized list or subject directory. Browse the appropriate category to make your selection. Or, you can use keywords, including natural language, to search for a database using the search engine. Searches should be kept simple.
Second, search the database you just located. Review the HELP screens for tips on improving the effectiveness of your search query.
According the same reference, [19], a selective list of Invisible Web Search Tools is:
Complete Planet (http://www.completeplanet.com/index.asp)
Direct Search (http://www.freepint.com/gary/direct.htm)
ProFusion (http://www.profusion.com)

FirstGov (http://www.firstgov.gov) -- access to federal government databases
Digital Librarian - http://www.digital-librarian.com
Infomine - http://infomine.ucr.edu/ - a virtual library of Internet resources relevant to faculty, students, and research staff at the university level.
MagPortal - http://magportal.com - Find individual articles from many freely accessible magazines.

3.2 Multiple Resources to Mine the Deep Web

The various deep web search resources can be classified such as: search engines, databases, catalogs, directories, social media and more, and guides.

Search engines
Whether you're looking for specific science research or business data, these search engines will point you in the right direction.

Databases
Tap into these databases to access government information, business data, demographics, and beyond.

Catalogs
If you're looking for something specific, but just don't know where to find it, these catalogs will offer some assistance.

Directories
Get hand-picked links to high quality research sources with these directories.

Social Media and More
Social media sites are a great way to find content that's obscure or hasn't quite made it to the search engines yet. Use these tools and more to round out your arsenal.

Guides
Use these guides to learn how to fine-tune your search on the invisible web. For example: “The Deep Web” **a** guide quickly discusses the deep web and offers a few tips for finding deep web information.

4 Case Study: Deep Web Search Tools

4.1 Webometrics for the Deep Web

There are many metrics for asses deep web searching. The main metrics proposed are from the following area [4]:

- PageRank
- Graph theory
- Network mapping
- Search engine optimization
- Impact factor

For example, one relatively straightforward measure is the "Web Impact Factor" (WIF) introduced by Ingwersen [2]. The WIF measure may be defined as the number of web pages in a web site receiving links from other web sites, divided by the number of web pages published in the site that are accessible to the crawler. However the use of WIF has been disregarded due to the mathematical artifacts derived from power law distributions of these variables.

4.2 Deep Web Search Tools

In order to be succesfull while searching the deep Web we provide you, learn how to use the three websites described below:
CompletePlanet™ uses a query based engine to index 70,000+ deep Web databases and surface Web sites. Appendix A lists 60 of the largest deep Web databases which contain 10% of the information in the deep Web, or 40 times the content of the entire surface Web. These 60 databases are included in CompletePlanet’s indexes. The interface is intuitive and easy to use. You can do a keyword search on all 70,000+ databases to find which databases to use for your search. You can also browse by category, and then search databases of interest.

ProFusion is a combination of query based engine and a deep Web directory portal. The directory structure is accessed by clicking on Specialized Searches. With an account, you can setup custom "My Search Groups" to search customized lists of websites and/or databases of your choice. For example, you could create a group called Technology and add all the databases and websites of interest to you. This group is saved to your profile. You could then, at any future time, search this group on a research topic with keywords. This is a great time saver. Their query based engine is called SmartDiscovery®.
SurfWax also uses a site's existing search capability as part of the meta-search process to tap the deep Web. They use proprietary algorithms to interpret the site's search criteria (Boolean, etc). With an account, you can also setup custom SearchSets to search customized lists of websites and/or databases of your choice. Surfwax also has a news accumulator feature with over 50,000 news topics in 84 categories. This news accumulator feature is a godsend providing high quality results. These are some useful news accumulator categories: all topics, networking, technology, telecommunication, and web services. In addition this site has WikiWax which takes the online encyclopedia Wikipedia to the next level. WikiWax does advanced look-aheads on Wikipedia searches to speed your keyword choices.

Finding Deep Web Resources

In addition to other methods discussed in this presentation, Schlein [10] shares several techniques to help the researcher find deep Web resources.

Pre-emptive search to find deep Web databases, use a search engine or search a site containing both surface and deep Web content. For example, to find a database containing information on viruses use this search term (exact syntax may vary among search engines):

- On Google or InfoMine search for virus *(database OR repository OR archive)* has this additional method specific for the Teoma search engine
- On Teoma search for: virus *(resources OR meta site OR portal OR pathfinder).*

Reverse-Link Searching: Find out which pages link to a database you already find useful and see if those sites have further recommendations. To do this, use the "link" operator in the search engine. For example, Google uses "link: yourURL." If you want to find out what sites link to NTIS, type this in the Google search bar: **link:http://www.ntis.gov**

Find Experts: When you do a search with Teoma, experts and enthusiasts for your keywords are listed to the right of the results column. Go to these sites and see what resources are recommended to help you "mine" for deep Web resources.

Search by document type: Search engines are now indexing heretofore "deep" files, like PDF files. In Google, by preceding your search terms with "filetype:ext" (where "ext" is the 3 character file extension), only those files will appear in the results. More **about Google:** When you do a search, the results are not only in the window you are viewing, but also simultaneously in the associated windows under the topics listed at the top of the search page, namely, *Web*, *Images*, *Groups*, *News*, *Froogle*, *Local*, etc. For example, if you search for the word "virus," under *Web* are the websites found for virus, under *Images* are the graphics found for virus, under *Groups* are the discussion groups on virus, etc. - all of this is available without you doing anything extra on your part other than click each topic link in succession.

The following tables illustrate some (not all) information situations where search engines are usually appropriate [17]:

Search For...	Choose
General topic	Google, Yahoo, MSN Search
Limited image search	Google, Yahoo, Picsearch, Ithaki, Ditto
Results clustered by subject for refining	Vivisimo, Clusty, Kartoo
Background information on an unfamiliar topic	Google, Yahoo, MSN Search
Find facts, calculate problems, convert formulas	Google, MSN Search
Multiple News Sources (current)	Google, Yahoo, World News Network
Country Information (Statistics, government,	CIA World Factbook, InfoNation

population, comparisons etc)	
Driving Directions, Maps	Google Maps, Mapquest, Mapmachine
Search by domain (.org, .com, .gov, .mil)	Google, Yahoo, MSN Search
Search for pages in specific languages	Google, Yahoo, MSN Search

Table 1. When to Use a usually Search Engine [17].

According to [17], searching into the invisible web is easier to make in the following conditions:

- Familiar with your topic
- Need a specific answer
- Need credible and verifiable information
- Need information contained in databases
- Information dynamically generated or changes frequently
- Topic is narrow
- Real-time information is needed (stock quotes, flight times, and other examples)
- Familiarity with specific research tools

Searching For...	**Choose Invisible Resources**
Is my flight on time?	Flight Arrivals, Yahoo Travel: Check Flight Status
What is the path of a hurricane?	National Hurricane Center
What concerts will be playing in Bucharest next month?	Pollstar, Ticketmaster
How do I find recent magazine articles for free on the web?	FindArticles, MagPortal
I would like to download sheet music of Bach's works.	Choral Public Domain Library

Table 2. When to Use the Invisible Web [17].

As an example of applying some of the principles in this presentation, let's do a search on "web mining" using a surface search engine and a Deep Web Database.

First let's do a surface search on Google. The result is hundreds million hits. I don't really have time to look at hundreds million hits; even a million might take a while. Realistically, I'll look at the first hundred or so and perhaps adjust my keywords then search again. The results show too many vendor sites and only a few dozen sites that might have good information. These results were above average for ten minutes of work. However, I will need to evaluate these few dozen sites and that could take a few hours – maybe I will find something useful among these, maybe not.

Now let's try a deep Web site like Educause, again using the same keywords, "web mining." There are round about 3,000 hits. I look at the first hundred or so and none of these are vendors. There are however many PDF files that look like they contain useful information.

Comparing the quality of the results between the two methods, for this search, the deep Web results have more substance and credibility. Of course this will not always be the case. The surface and deep Web each have their advantages and disadvantages depending on the search topic. You need both aspects of the Web plus a phone to call people (not all information is on the Web). In this example, within three minutes, the deep Web search revealed a goldmine of high-quality information very relevant to the search topic.

4.3 Search Engine Inconveniances

People's access to content on the Web has been greatly enhanced by search engines also, these search engines make a research work succesful and they are an important component of the research procedure. In table 3 are presented a list of the limits for work with search engines [19].

Inconveniences	**Details**
Search Engine Display Limit	The results displayed by the search engines from a specific site may have a limit. For instance, only one or

	two results, may be displayed in the search results for each site Google indexes. But, unless you click on more results other relevant pages from that site will not be displayed even though they might be indexed.
Quality	It is not considered when sites whose content might not be accurately reflected by their name will appear.
Quantity	SE returns tens of thousands of results with no way of separating the quality sites from the useless ones.
Search Engine Preferences	Depending on user preferences, a search engine might be able to find available documents but the user preferences are screening it out.
Surface Indexing	A search engine might index the home page and maybe a page or two after that from a large site. It will not return all the content from a quality web site with many levels of information.
Business & Popularity	Most search engines rank pages according to popularity. Google, in particular, returns results based on the most popular and best-known sites for your topic (how many pages linking to a site are a big part of Google's results). Business organizations pay big bucks to have their sites appear near the top of a search. Many search engines target consumers looking to buy products. Information seekers have become secondary targets for some search engine companies.

Table 3. Search engine inconveniences.

Conclusion

The search tools for exploring the Invisible Web need the new valuable resources because the usual search engines can't be available to enlighten searchers. With enhanced resources for the hidden contents of the deep web documents the search process is more efficient and productive.
We can get benefits beyond just academic, by knowing the Invisible Web. Through the Invisible Web we can gain information concerning medicine, law, and other areas. Also, people can become informed consumers. One of the benefits of mining to use the invisible web resources in conjunction with search engines is how to be able to discern viable information from information targeting consumers.
In our case study we emphases multiple resources; which is, for now, an accepted solution for exploring the Invisible Web. Our contributions from this case study and the trials we made by using the "Federated Search" techniques in searching the Invisible Web further on, encourage us to improve new tools for mining the Deep Web without using up too many resources.

References

[1] S. Brin, L. Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, http://infolab.stanford.edu/~backrub/google.html
[2] P. Ingwersen, *The calculation of web impact factors*, Journal of Documentation, Vol. 54, No. 2, March 1998.
[3] E. T. Peterson, *Website Measurement Hacks,* Published by O'Reilly & Associates, 2005.
[4] I. Pop, *Web Metrics using Web Quality Model in Social Media*, New Trends in Approximation, Optimization and Classification, Proceedings of International Workshop, Sibiu, 2008, ISBN: 978-973-739-678-5 pp. 58-67, 2008.

[5] V. Shkapenyuk, T. Suel, *Design and Implementation of a High-Performance Distributed Web Crawler*, Technical Report TR-CIS-2001-03, CIS Department Polytechnic University Brooklyn, NY 11201, 2001.
[6] J. Sterne, *Web Metrics: Proven Methods for Measuring Website Success*, Published by John Wiley & Sons. Inc, 2002.
[7] P. Tin, T. T. Zin, H. Hama, *A Novel Interdisciplinary Approach to Deep Web Search Engine*, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008
[8] W. Warnick, *Searching the Deep Web*, D-Lib Magazine January 2001, Vol. 7 N# 1, 2001.
[9] http://www.philb.com/webse.htm
[10] http://techdeepweb.com/4.html
[11] http://www.webometrics.info/methodology.html
[12] http://www.uaf.edu/Library/instruction/handouts/Invisible_Web.html
[12] http://netforbeginners.about.com/library/diagrams/n4layers.htm
[13] http://en.wikipedia.org/wiki/Deep_Web
[14] http://www.searchengineshowdown.com/
[15] http://searchenginewatch.com/.
[16] http://www.websearchguide.ca/research/
[17] http://eastlrc.valencia.cc.fl.us/invisibleweb/comparing_search_engines_to_the.htm/
[18] http://www.webdesignerforum.co.uk/index.php?autocom=blog&blogid=85&showentry=153/
[19] http://www.valenciacc.edu/library/east/invisible_reasons.cfm

IOAN POP
"Lucian Blaga" University - Sibiu
Informatics Department
Dr. Ioan Rațiu street, nr. 5-7
ROMANIA
E-mail: me.ioanpop@gmail.com