INVISIBLE WEB AND KNOWLEDGE DISCOVERY TOOLS: A STUDY

G RATHINASABAPATHY

Abstract

Invisible Web which is also known as 'Deep Web' or 'Hidden Web' refer to information content that is 'invisible' to conventional search engines. Public information on the invisible web is currently around 500 times larger than the commonly defined World Wide Web. Nearly 550 billion individual documents are available in the invisible web while the surface web has around only one billion documents. It has been estimated that around 2,00,000 invisible web sites exist on the Information Superhighway at present and is the largest growing category of new information on the Internet. The total quality content of the invisible web is also greater than that of the surface web and highly relevant to every information need, market and domain. But, a larger portion of the invisible web is missing from search engines' results pages. In this context, this paper attempts to characterize the deep web's content and relevance to information seekers and profile special tools available to mine the invisible web.

Keywords: Invisible Web/ Hidden Web/ Deep Web/ Knowledge Discovery Tool/ Search Engines

1. Introduction

Internet content is considerably more diverse and the volume certainly much larger than commonly understood. In the early days of the Internet, it was reasonably easy to find information or data files using a variety of software that were usually command driven. However, with the proliferation of data brought about by the growth of the web, the systems such as Archie, Gopher and Veronica became increasingly unable to cope. In order to overcome the lack of retrieval facilities, a number of search engines came which prove to be of great assistance in allowing the information seeker to quickly find the piece of data that they require. But, the search engines all have their own shortcomings. The major shortcoming is the widely used search engines which are known as 'general purpose search engines' do not index most of the contents forming the invisible web.

2. Invisible Web

'Invisible Web' is a phrase first coined by Dr. Jill Ellsworth in 1994 to refer to information content that was 'invisible' to conventional search engines. Invisible web is also known as 'Deep Web' or 'Hidden Web'. A large part of the invisible web is comprised of databases that only reveal their information if search request is made directly. General purpose search engines do not find this dynamically-generated information. The primary information stored on the invisible web are usually various non-textual file formats and free content-rich databases created by Government agencies, educational institutions, and other organizations around the world [1]. Phone books, people finders, patents, laws, dictionaries, digital exhibits and multimedia geographical files are some of the examples of the contents available on the invisible web.

Since the invisible web contains information that is new and dynamically changing in content, researchers have taken a lot of efforts to design search tools to retrieve valuable contents from the Invisible web. Those search engines crawlers and indexing programs have overcome many of the technical barriers that made it impossible for them to find and provide invisible web pages.

Previously pages in non-HTML formats were excluded by the search engines. But, now most of the search engines translate all those non-HTML files such as pdf, Word, Excel, Corell suite, etc. into HTML and provide them in the search results. Similarly, script-based pages, whose links contain a ? or other script coding, no longer cause most search engines to exclude them. Pages generated dynamically by other types of database software (e.g., Active Server Pages, Cold Fusion) can be indexed if there is a stable URL somewhere that search engine spiders can find. There are now many types of dynamically generated pages like these that are found in most general web search engines.

There are still some hurdles search engine spiders cannot leap, and these still create a huge set of web pages not found in general search engines because search engines still cannot type or think of its own. If access to a web pages requires typing, web crawlers encounter a barrier they cannot go beyond and they cannot search online catalogues and they cannot enter a password or login. Most of the invisible or deep web is made up of the contents of thousands of specialized searchable databases made available via the web. When we type a search in one of these databases, the search results are delivered to us in web pages that are generated just in answer to our search.

2.1 Invisible Web: Some Basic Facts

The term 'invisible web' mainly refers to the vast repository of information that search engines and directories don't have direct access to, like databases. The size and relevancy of the invisible web has been quantified as follows [2]:

- Around 95 per cent of the invisible web is publicly accessible and not subject to fees or subscription.
- Contains 7,500 terabytes of information compared to nineteen terabyes of information in the surface web.
- It is growing at a much faster rate than the surface web.
- Many experts believe that the total quality content of the invisible web is greater than that of the surface web.
- More than 2,00,000 invisible web sites presently exist on the Information Superhighway.
- More than half of the invisible web content resides in topic-specific databases.
- Most of the information on the invisible web is maintained by academic institutions and research organizations.
- Nearly 550 billion individual documents are available at present in the invisible web compared to the one billion of the surface web.
- Public information on the invisible web is currently 500 times bigger than the searchable or surface web.
- The content is highly relevant to every information requirement, market and domain.

2.2 Importance of Invisible Web

It is not always easy to find the information with a search engine, especially if we are looking for something a bit complicate or obscure. This is where search engines will not necessarily help us, and the invisible web will. Further, the search engines only search a very small portion of the web which makes the invisible web a very tempting resource. It has been reported that 85 per cent of web users use search engines to find needed information, but nearly as high a percentage cite the inability to find desired information as one of their biggest frustrations.

2.3 Subject Coverage of Invisible Web

A study undertaken to know the subject coverage across 17,000 invisible web sites revealed that a surprisingly uniform distribution of content across all areas, with no category lacking significant representation of content [3]. It is clear from Table-1 that invisible web content also has relevance to every information need and market.

Table 1: Subject Coverage of Invisible Web (in %)

Agriculture	2.7	Law / Politics	3.9
Arts	6.6	Lifestyles	4
Business	5.9	News / Media	12.2
Computing / Web	6.9	People / Companies	4.9
Education	4.3	Recreation / Sports	3.5
Employment	4.1	References	4.5
Engineering	3.1	Science / Mathematics	4
Government	3.9	Travel	3.4
Health	5.5	Shopping	3.2
Humanities	13.5		
Health	5.5	Shopping	3.2
Humanities	13.5		

2.4 Content Coverage of Invisible Web

The major information resources stored in the invisible web are normally in non-textual formats and free content-rich databases created by Government agencies, educational institutions and other organizations around the world. They include patents, digital exhibits, laws, dictionaries, phone books, people finders, items in web stores or auctions, multimedia and geographical files. Further, the information is usually new and dynamically changing in content such as news, job postings, flight schedules, accommodation reservation, stock prices, etc. [4]

3. Invisible Web Vs. Search Engines

Most of the general purpose search engines claim that they 'index the web'. But they are actually indexing a very small portion of it. These search engines using 'spiders' and 'crawlers', were designed to index simple HTML pages that have incoming links from other pages on the web. But modern web sites, operating databases to generate pages on-the-fly, are 'too sophisticated' for these search engines to index their pages. So while these search engines do a very good job at indexing small sites and personal home pages, they cannot provide sophisticated sites and databases. Therefore, these search engines which are unable to index most of the web are providing the user with a partial service. For example, we can take Google as an example. It is considered by most of the people that it has about eight billion pages in its index. Those eight billion pages seem like a lot until consider that the Deep Web is estimated to be 500 times bigger than the searchable Web. If we multiply 500 by the 8 billion in Google's index we can understand the fact that Google is only indexing a fraction of the searchable Web.

4. Invisible Web Search Tools

Realising the importance of the problems due to invisible web, several organizations started offering solutions which includes search engine exposure technologies for large e-commerce and content sites, and cutting-edge search, retrieval, and information extraction platform for exposing hidden information on the cyberspace. A prototype invisible web crawler was built at Stanford to crawl the invisible web [5]. The researchers have introduced a new Lay out based Information Extraction Technique (LITE) and demonstrated its use in automatically extracting semantic information from search forms and response pages.

The Deep Web Technologies, an expert at mining the invisible web, built ExploritTM a proprietary software platform that quickly and efficiently locate and deliver 'difficult to find' information. The software which has powerful and comprehensive web-based interfaces that consolidate key sources permit real-time information search, retrieval, aggregation, and relevance ranking and facilitate access to the 94 per cent of the planet's electronically-stored data and documents that is missed by common consumer search engines such as Google [6]. A number of similar proprietary software platforms are available to mine the invisible web.

4.1 Invisible Web Search Engines

Though vast expanses of the web are invisible to general purpose search engines, the following search engines furnished in Table-2 will help to search the invisible web or deep web. It is only a select list of invisible search engines and few more such search engines are also available on the cyberspace.

Table 2: Invisible Web Search Engines

Name of Web Directory	Uniform Resource Locator
Invisible Web	www.invisible-web.net
LexiBot	www.lexibot.com
WebData	www.webdata.com
Turbo10	http://turbo10.com
CompletePlanet	www.completeplanet.com/index.jsp

4.2 Web Directories

Invisible web can be mined through using the web directories which are also known as web indices. Web directories are different from search engines. These are typical directories offer search capability, but it is essentially a hierarchical list of web sites, organized into categories. Each category has others below it, with contents listed alphabetically. Unlike search engines, directories are hand-compiled by human beings like a library catalogue. A select list of web directories is furnished in Table-3.

Table 3: Web Directory

Name of Web Directory	**Uniform Resource Locator**
Clearing House	www.clearinghouse.net
Digital Librarian	www.digital-librarian.com
GeniusFind	www.geniusfind.com
IncyWincy Search	www.incywincysearch.com
Infomine	http://infomine.ucr.edu
InvisibleWeb.com	www.invisibleweb.com
Librarians Index to the Internet	www.lii.org
Allestra	www.allestra.com
DMOZ (Open Directory)	www.dmoz.org
With1Click	www.with1click.com
IcySpicy	www.icyspicy.com/web.html
4Anything	www.4anything.com
Zeal www.zeal.com	
Gimpsy	www.gimpsy.com
MavicaNet	www.mavicanet.com
Hoppa	http://hoppa.com
JumpCity	www.jumpcity.com
SunSteam	www.sunsteam.com
Shadowood	www.shadowood.com
NoSearch	www.nosearch.com

Scholarly and academic research contents are also not easily searched by the general purpose search engines because they are great for searching business and popular information alone. Therefore, they are not very useful for finding scholarly and academic research contents or scholarly journal articles. Journal articles are available in a variety of formats, ranging from citations or brief abstracts to full-text delivered electronically and a huge collection of such scholarly contents are provided free. These invaluable resources abound online can be found using the tools furnished in Table-4 below.

Table 4: Tools to mine Scholarly Contents

Name of Web Directory	**Uniform Resource Locator**
Name of Web Directory	Uniform Resource Locator
AllAcademic	www.allacademic.com
Eric Database	www.eric.ed.gov
FindArticles.com	www.findarticles.com
Google scholar	http://scholar.google.com

HighBeam	www.highbeam.com
InfoTrieve	www4.infotrieve.com
JSTOR	www.jstor.com
MagPortal	www.magportal.com
Periodicals.Net	www.periodicals.net
SciBase	www.thescientificworld.com/DBProducts/ sciBASE
SearchEbooks	www.searchebooks.com
SearchEdu	www.searchedu.com
Silver Platter	www.silverplatter.com
To Find / Search Theses and Dissertations	
Digital Dissertations	Wwwlib.umi.com/dissertation/
Dissertations.com	www.dissertation.com
DissertationsAndTheses.com	www.dissertationsandthesis.com
Theses.Org	www.theses.org

4.3 Virtual Libraries

The Virtual Library is the oldest catalogue of the web, started by Tim Berners-Lee, the father of the World Wide Web and HTML. They are also known as gateways, digital collections, digital libraries and cyber libraries. The Virtual Library is not a commercial catalogue. It is run by a loose confederation of volunteers, who compile pages of key links for particular areas in which they are expert. Even though it is not the biggest index of the web, the virtual library pages are widely recognized as being amongst the highest quality guides to particular sections of the web. The virtual library is not living in one place. It is a collection of web pages and indexes live on hundreds of different servers around the world. A set of catalogue pages linking these pages is normally maintained at the virtual library site. Each maintainer is responsible for the content of their own pages. Information is checked for accuracy and authority before linking them to the virtual libraries. The data made available in virtual libraries are displayed clearly and concisely, allowing for easy navigation and mostly they are current [7]. A detailed list of virtual libraries available on various subjects is available at http:// vlib.org

5. Conclusion

The invisible web is quickly becoming a very important place to those seriously looking for quality information on the Information Superhighway. Since most of the information on the invisible web is maintained by academic institutions, and has a higher quality than search engine results, there is a great demand for these contents. Therefore, it is absolutely necessary that the information professionals need to connect the information seeking community who have been using only general purpose search engines to the quality and rich content of the invisible web. Further, the search tools presently available to mine the invisible web need to be improved to reveal more invisible and hidden contents since so many searchers depend on the free web search tools for their information needs.

References

1. Lackie, Robert J. (2003). "The Evolving "Invisible Web": Tried-and-True Methods and New Developments for Locating the Web's Hidden Content". College and Undergraduate Libraries 10(2): 65-71.
2. Wendy Boswell. "The Invisible Web". Available at < http://websearch.about.com/od/invisibleweb/a/invisible_web.htm> accessed on 30-12-2006.
3. Bergman, Michael, K. (2001). "The Deep Web: Surfacing Hidden Value". The Journal of Electronic Publishing 7(1):
4. Cohen, Laura (2002). "The Deep Web". University at Albany Libraries: Internet Tutorials. Available at <http://library.albany.edu/interent/deepweb.html>Accessed on 15-12-2006
5. Sriram Raghavan and Hector Garcia-Molina (2001). "Crawling the Hidden Web". In Proceedings of the 27th Very Large Database Conference, Roma, Italy, 2001. 129-138.
6. Deep Web Technologies available at <http://www.deepwebtech.com> accessed on 18-12-2006.
7. Rathinasabapathy, G. (2006). "Virtual Library on Poultry Diseases: Design, Development and Evaluation". In Compendium of SALIS 2006 Conference on Initiatives in Libraries and Information Centres in the Digital Era, Coimbtore, India. 70-75.

BIOGRAPHY OF AUTHOR

Dr. G. Rathinasabapathy is currently working as Assistant Librarian (S.S.) in Tamilnadu Veterinary and Animal Sciences University, Chennai. He holds M.Com., M.L.I.S., M.Phil. (LIS) and Ph.D. (LIS). He has over 12 years of professional experience in the field of Library and Information Science. He has published five books, 15 scientific papers and 120 articles. Participated and presented papers in various national and international Seminars, Symposia and Conventions.

Email : grspathy@yahoo.com