DEEP WEB DIRECTORY: A SPECIAL REFERENCE TO COMPLETE PLANET

[1]Dr. S. Srinivasaragavan[1] M. Surulinathi[2] B. Neelakandan[3] R. Balasubramani[4]

Abstract

As the web grows, content providers will find themselves with the difficult task of disseminating information. Presently, users depend on the popular search engines and portals for their outlet for information. Unfortunately, these search tools don't have the capabilities to gain access to databases which hold valuable data. To have access to these databases, one must become familiar with the structure of the invisible web. CompletePlanet is the front door to these Deep Web databases on the Web and to the thousands of regular search engines, it is the first step in trying to find highly topical information. By tracing through CompletePlanet's subject structure or searching Deep Web sites, you can go to various topic areas, such as energy or agriculture or food or medicine, and find rich content sites not accessible using conventional search engines. By nature, this listing is preliminary and likely incomplete, since we lack a complete census of Deep-Web sites. This inability today to identify all of the largest Deep-Web sites should not be surprising. The awareness of the Deep-Web is a new phenomenon and has received little attention.

Introduction

There are hundreds of thousands of databases that contain Deep Web content. CompletePlanet is the front door to these Deep Web databases on the Web and to the thousands of regular search engines, it is the first step in trying to find highly topical information. By tracing through CompletePlanet's subject structure or searching Deep Web sites, you can go to various topic areas, such as energy or agriculture or food or medicine, and find rich content sites not accessible using conventional search engines. By nature, this listing is preliminary and likely incomplete, since we lack a complete census of Deep-Web sites. This inability today to identify all of the largest Deep-Web sites should not be surprising. The awareness of the Deep-Web is a new phenomenon and has received little attention.

What is the Deep Web?

The Deep Web is content that resides in searchable databases, the results from which can only be discovered by a direct query. Without the directed query, the database does not publish the result. When queried, Deep Web sites post their results as dynamic Web pages in real-time. Though these dynamic pages have a unique URL address that allows them to be retrieved again later, they are not persistent

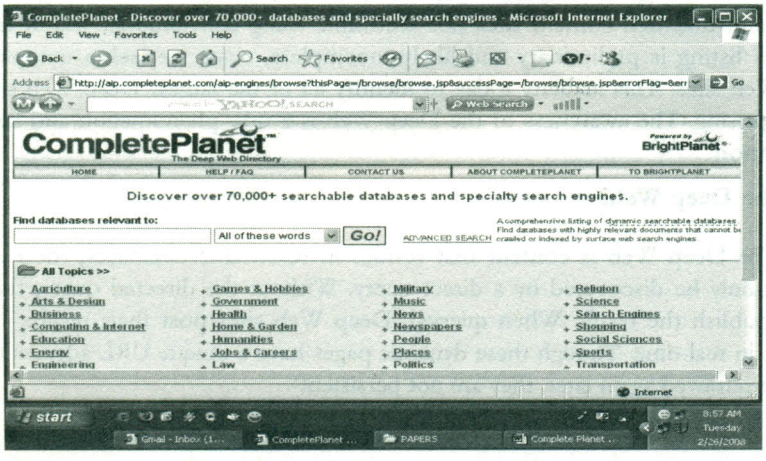
[1]*Librarian & Head, Bharthidasan University, Trichy*
[2]*Guest Lecturer, Library Assistant (SG,)DLIS, Bharathidasan University,Tiruchirappalli*
[3]*Deputy Librarian, Kumaraguru College of Technology, Coimbatore, surulinathi@gmail.com*

How to find "Invisible Web" resources?

- ProFusion, http://www.profusion.com
- CompletePlanet: Discover and Search 103,000 Searchable Databases and Specialty Search Engines, http://www.completeplanet.com/
- Gary Price's Direct Search, http://www.freepint.com/gary/direct.htm
- Librarians' Index to the Internet, http://www.lii.org
- Resource Discovery Network - http://www.rdn.ac.uk/
- "as close to a search engine for the hidden web as you're likely to find," SearchDay, 4 Dec. 2001 (http://searchenginewatch.com/searchday/01/sd1204-rdn.html)
- Invisible-Web Net, http://www.invisible-web.net/
- The Big Hub - http://www.thebighub.com/
- searchable; also includes directory listing of over 3,000 subject specific searchable databases
- Turbo 10, http://turbo10.com/

How does the Deep Web differ from the "surface" Web?

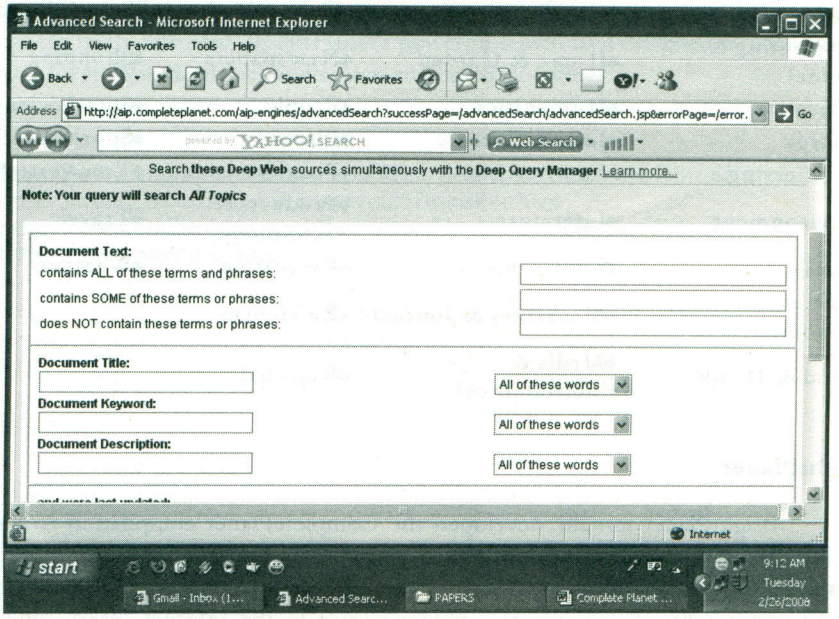
Search engines — the primary means for finding information on the "surface" Web — obtain their listings in two ways. Authors may submit their own Web pages for listing, generally a minor contributor to total listings. Or, search engines "crawl" or "spider" documents by following one hypertext link to another. Simply stated, when indexing a given document or page, if the crawler encounters a hypertext link on that page to another document, it records that incidence and schedules that new page for later crawling. Like ripples propagating across a pond, in this manner search engine crawlers are able to extend their indexes further and further from their starting points. Thus, to be discovered, "surface" Web pages must be static and linked to other pages. Traditional search engines cannot "see" or retrieve content in the Deep Web, which by definition is dynamic content served up in real time from a database in response to a direct query.



http://aip.completeplanet.com/aip-engines/browse?thisPage=/browse/browse.jsp&successPage=/browse/b

Advance Search

Advance search is designed for filter, straightforward searches. The interface allows you to search on a variety of fields from pull-down menus, including phrase, Boolean queries and more. You can then select Boolean operators from drop-down menus between the search boxes.



Subject Coverage:

Collectively, the value of these databases is extremely high. Each database is very focused in nature and the sheer numbers of them indicate that there are hundreds if not thousands in any given subject area. If you're just surfing the Web, you can click on the link or links provided by CompletePlanet, go to the individual high value databases, and search one-by-one there. But, if you need to tap into many of them in a given subject area, you need more powerful tools to find information you can't get by any other means. Such a tool already exists and is called the Deep Query Manager (DQM).

DQM is the kind of access point that is highly valuable to professional searchers and knowledge professionals. When you're looking to shine a bright light in all the deep dark crevices and corners of the Web and make sure you can comprehensively discover all the content publicly available, this is the resource you need.

It really is a generations long challenge: to identify, organize and serve up the most relevant results from what is now the global information repository. BrightPlanet's technologies are specifically designed to address those needs. Its services are built around a complete set of automated tools for identifying, talking to, retrieving, qualifying, placing and publishing the quality documents that can be obtained from those sites. It covers the following disciplines.

✦Agriculture	✦Games & Hobbies	✦Military	✦Religion
✦Arts & Design	✦Government	✦Music	✦Science
✦Business	✦Health	✦News	✦Search Engines
✦Computing & Internet	✦Home & Garden	✦Newspapers	✦Shopping
✦Education	✦Humanities	✦People	✦Social Sciences
✦Energy	✦Jobs & Careers	✦Places	✦Sports
✦Engineering	✦Law	✦Politics	✦Transportation
✦Environment	✦Literature	✦Products & Technology	✦Travel
✦Family	✦Living things	✦Recreation	✦Weather
✦Finance & Economics	✦Magazines & Journals	✦References	
✦Food & Drink	✦Media & Entertainment	✦Regional	

BrightPlanet

BrightPlanet initially developed the CompletePlanet compilation to identify and tap into many hundreds and thousands of search sources simultaneously to automatically deliver high-quality content to its corporate and enterprise customers. It then decided to make CompletePlanet available as a public service to the Internet search public.

http://brightplanet.com/

The challenge is how to harness massive amounts of data in poorly-used document assets, from within the organization and globally. The dimensions of the "content challenge" are truly daunting. Paper and electronic documents in huge numbers reside in many different forms and stovepiped locations. To complicate matters there are different formats, media, languages and electronic encodings, and different semantics and meanings. Like so many scattered nuggets of gold, these valuable assets are everywhere.. but now are ready to be put to productive use. Search, harvest, consolidate, index, merge, analyze and categorize documents and associated metadata:

- In any format
- In any language
- From surface *and* deep sources
- From inside *and* outside the firewall.

Conclusion

As the web grows, content providers will find themselves with the difficult task of disseminating information. The essential conclusion is that time is well-spent in understanding how to pose a proper query and how to take advantage of the way that search services work. Presently, users depend on the popular search engines and portals for their outlet for information. Unfortunately, these search tools don't have the capabilities to gain access to databases which hold valuable data. To have access to these databases, one must become familiar with the structure of the invisible web. The invisible web holds the majority of the documents that are on the internet and are rarely retrieved. BrightPlanet's powerful search tool designed to harvest from thousands of Deep Web databases and search engines at one time.

References

1. http://www.brightplanet.com/resources/details/tutorial-part-2.html#topic2
2. http://www.powerhomebiz.com/vol25/invisible.htm
3. http://www.brightplanet.com/resources/details/tutorial-part-7.html
4. http://www.brightplanet.com/resources/details/deepweb.html