

Search Engines Crawl Deeper

By Paul J. Bruemmer

Search Engines Crawl Deeper

Paul J. Bruemmer
paul2@web-ignite.com

Web Ignite Corporation <http://www.web-ignite.com>

Until recently, you couldn't access all the content in the World Wide Web because dynamic databases and many file formats were out of reach for Web crawlers. This is changing, as the Web morphs into a wider collection of data with audio and video files, as well as PDF, Excel, Power Point and more.

While the "surface Web" contained only text documents, major search engines now have the technology to index the "deep Web," a vast depository of previously untapped content in dynamic databases.

The difference between the surface and deep Web is both qualitative and quantitative. Qualitatively, deep Web content includes images, sounds, presentations, and many types of media invisible to search engine crawlers. Quantitatively, it was estimated to be about 500 times larger than the surface Web, although this can be misleading.

How Deep?

A BrightPlanet study conducted in March 2000 estimated public information in the deep Web to be about 500 times larger than the existing World Wide Web. The study stated the deep Web "contains billions of documents in hundreds of thousands of specialty databases hidden from public view." This may be academically correct, but it might not be useful to index all of these documents.

Some estimates say the useful deep Web is merely two to three times the size of the surface Web. That's because some dynamic sites can generate millions of variations of the same page with content management solutions that personalize pages on price, size, etc. Should each of these be considered a unique page?

Dynamic Site Indexing

Most major search engines (AltaVista, FAST, Google, Inktomi, Lycos, etc.) now index dynamic content. Thanks to paid-inclusion programs, search engines will index additional dynamic content for a fee. Some of these programs include AltaVista Trusted Feed, FAST PartnerSite, Lycos InSite, and Inktomi Search/Submit.

These premium services do not guarantee positioning like the pay-per-click programs (Overture, FindWhat, etc.). Rather, they will index dynamic content and include more frequent refreshes. While pricey, it's ideal for submitting Web pages traditionally difficult to crawl (large database dynamic sites and framed sites).

Deep Web Search Tools

There are a number of products and services that enhance deep Web searching, including BrightPlanet, Intellisearch's Invisible Web, ProFusion, Quigo, and C|Net's Search.com.

One of these, Quigo, is capable of retrieving, normalizing and indexing documents in an offline crawling process, which enables users to submit queries to thousands of sites at once. The others named above focus on expanding the meta-search engine concept, returning queries from about a dozen sources.

Quigo technology operates behind the scenes, allowing portals and search engines to access and manage dynamic Web content not currently indexed by traditional search engines. It maps these pages and uses Information Extraction (IE) algorithms to restructure the information within a page, keeping each piece of data within its relevant context. Restructuring works as follows:

1. Categorized results – Upon each query, users are presented with a list of all relevant categories in which results were found. This enables quick refinement, pinpointing the most relevant information (a search for "ford" presents categories such as person, car, actor, company, etc.).
2. Associative search – Certain attributes are hyperlinked in each search result. By clicking the linked words, users can traverse within Quigo's deep Web database to locate other similar documents (a search for "Jurassic Park" will bring up several book sites; click the author's name, and all books found by Michael Crichton are displayed).

This ability to restructure data can be expanded in many ways. Information can be sorted by date, by price in comparison shopping, or it can be used to locate the best performing stocks, to find the closest coffee shop by zip code, and so forth. Since deep Web sites can be among the most authoritative sites on the Web, this solution can help overcome the relevancy issue, providing a highly useful service for users as well.

It's apparent that Web search continues to change as the Web matures. The portals, engines, and directories that give users the best results are those that will prosper and help define the nature of Web search in the days to come.

Bio

Paul J. Bruemmer is the CEO of Web Ignite, <http://www.web-ignite.com/> a professional search engine marketing company. Founded in 1995, Web-Ignite has helped promote over 15,000 Web sites. Client testimonials report traffic increases of 150 to 500 percent. Bruemmer presents at search engine conferences and his articles have appeared on ClickZ, New Media, Internet Day, B2B Interactive, I-Advertising, MarketingProfs, Marketing Sherpa, Search Engine Guide and SitePoint.

[Get-Articles.com : 1000's of reprintable business and internet marketing-related articles.](#)

[Submit your article for reprint.](#)