Congreso Internacional INFO'2004

Para acceder al web profundo: conceptos y herramientas

Autores: ***Lourdes Vilaragut Llanes***
Juan R. Carro Suárez

Institución: Consultoría Biomundi / IDICT
Calle 200 #1922 e/ 19 y21. Atabey. Playa. Ciudad de La Habana. Cuba
carro@biomundi.inf.cu; lourdes@biomundi.inf.cu

RESUMEN

Hoy buscar información en Internet puede llegar a convertirse en una experiencia poco agradable, puede recuperarse gran cantidad de información irrelevante y no encontrar lo que necesita. No sólo porque hay que saber cómo utilizar los llamados buscadores o máquinas de búsquedas tradicionales para obtener el máximo provecho, sino porque éstos apenas indizan una pequeña parte de todo lo que la red puede ofrecer. Buscar información utilizando los buscadores tradicionales puede ser comparado con arrastrar una red en la superficie de un océano. No podrá obtener muchos peces de aguas profundas.

La empresa BrightPlanet sostiene, merced a un estudio basado en datos recogidos entre el 13 y el 30 de marzo de 2000, que la disponibilidad de información pública en el llamado *Deep Web* o Web Profundo es alrededor de 550 veces mayor que en el Web Superficial *(Surface Web)*, lo que comúnmente llamamos *World Wide Web*.

El Web Profundo almacena páginas dinámicas que se obtienen en respuesta a interrogaciones directas a bases de datos; almacena documentos en formatos específicos diferentes de html, como por ejemplo pdf, doc, software, audio, videos, entre otros. La mayor parte de esta información no se recupera utilizando los buscadores tradicionales.

En este trabajo se pretende dar a conocer qué es el *Deep Web* o Web Profundo y mostrar algunas de las herramientas que existen en la actualidad para acceder a la información que en él se encuentra.

Palabras claves : web profundo, buscadores, internet

Para acceder al web profundo: conceptos y herramientas

Autores: ***Lourdes Vilaragut Llanes***
Juan R. Carro Suárez

La mayor parte de los usuarios de internet, cuando necesita buscar información, lo hace a través de las máquinas de búsquedas o directorios tradicionales, a los cuales se les llamara en esta ponencia simplemente buscadores. Esta búsqueda puede llegar a convertirse en una experiencia poco agradable porque puede recuperarse gran cantidad de información irrelevante y no encontrar lo que se necesita. No sólo porque hay que saber cómo utilizar los buscadores para obtener el máximo provecho, sino porque éstos apenas indizan una pequeña parte de todo lo que la red puede ofrecer.

Cada máquina de búsquedas utiliza su propio mecanismo de araña (robot o spider o rastreador) para recorrer la red siguiendo los enlaces o hipervínculos que se encuentran en las páginas estáticas por donde pasan, conformando una base de datos con la información recuperada. Los caminos seguidos por los distintos buscadores no son exactamente los mismos. A pesar de su pretendida exhaustividad, se calcula que las mayores máquinas de búsqueda (Google, AlltheWeb) indizan sólo un 16% de toda la información contenida en internet. Según estadísticas realizadas por sitios de reconocido prestigio internacional, se puede conocer que existe poco solapamiento en las bases de datos de los buscadores, lo cual tal vez justifica su proliferación, ya que cada uno va cubriendo diferentes áreas del espacio web, sin que por ahora sea posible técnicamente que ninguno sea exhaustivo.

Los buscadores arrojan resultados sobre las búsquedas realizadas en sus propias bases de datos y no sobre toda la web directamente.

La parte de la web que es accesible a través de los buscadores tradicionales se le conoce con el nombre de Web Superficial *(Surface Web)*. De esta parte de internet se escapan muchas veces cientos de miles de bases de datos importantes, entre las cuales se encuentran catálogos de bibliotecas, bases de datos bibliográficas, revistas electrónicas en las que es necesario un registro previo, obras de referencia como enciclopedias, diccionarios y otras. Esta información sólo puede ser accedida a través de búsquedas directas a los sitios donde están almacenadas, que pueden tener sistemas de búsqueda y recuperación propios y que como respuesta pueden conformar páginas que son creadas dinámicamente.

La parte de la web formada estos sitios y fuentes de información se le conoce con el nombre de Web Profundo o *Deep Web.* Algunas personas lo llaman, erróneamente, Web Invisible. Al hablar del Web Profundo algunos se refieren a bases de datos especializadas, archivos en formatos no html, como son pdf, doc, archivos de audio, video, imágenes, así como también bibliotecas virtuales, bibliotecas digitales y otros repositorios de información.

En 1994 el Dr. Jill Ellsworth fue el primero en acuñar la frase Web Invisible para referirse a la información que fuera invisible a las máquinas de búsquedas o directorios tradicionales. El término Web Invisible se dice que es inexacto porque:

- Muchos usuarios asumen que la única forma de acceder a la *web* es consultando un buscador.
- Alguna información puede ser encontrada más fácilmente que otra, pero ésto no quiere decir que esté invisible.
- La *web* contiene información de diversos tipos que es almacenada y recuperada en diferentes formas.
- El contenido indizado por los buscadores de la *web* es almacenado también en bases de datos y disponible solamente a través de las interrogaciones del usuario, por tanto no es correcto decir que la información almacenada en bases de datos es invisible.

Según estudios realizados por la compañía BrightPlanet (comenzó a hablar del Web Profundo en el año 2000) se dice que:

- La información pública del Web Profundo es actualmente alrededor de 550 veces mayor que la del Web Superficial.
- El Web Profundo contiene 7,500 terabytes de información comparado con 19 terabytes en el Web Superficial.
- El Web Profundo tiene el mayor crecimiento de nueva información en Internet.
- Más de la mitad del contenido del Web Profundo reside en bases de datos específicas.
- La calidad del contenido del Web Profundo es considerada por lo menos 1,000 o 2,000 veces mayor que la del Web Superficial.
- El 95% de la información del Web Profundo es información totalmente pública, libre de subscripciones y tarifas.

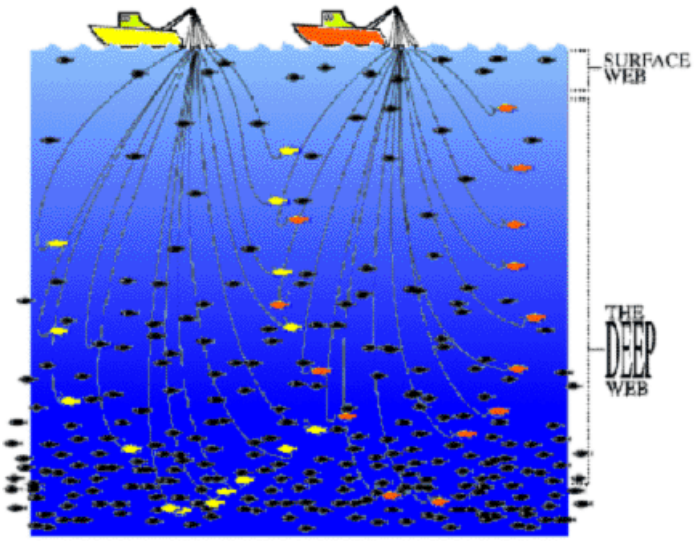


Gráfico de BrightPlanet. El Web Profundo y el Surface Web

En el Web Profundo o *Deep Web* puede encontrarse información que es válida para sistematizar en una base de datos, ej: directorios telefónicos, patentes, leyes, diccionarios, archivos gráficos y multimedia; información que es nueva y cambia continuamente su contenido, ej: noticias, avisos de trabajos, programación de viajes por avión o transporte terrestre, habitaciones libres en hoteles, información de los

mercados, clasificados, etc. También pueden encontrarse sitios de compañías, empresas, instituciones, etc.; páginas blancas y páginas amarillas; páginas internas de sitios muy grandes que son creadas dinámicamente. ej: base de conocimientos en el sitio de Microsoft.

Buscadores del Web Profundo

Estas herramientas permiten acceder a una mayor porción del web ya que, además de buscar en el Web Superficial, buscan en el Web Profundo, que resulta inaccesible para los buscadores tradicionales, en su mayor parte.

CompletePlanet: www.completeplanet.com
Pertenece a la compañía BrightPlanet. Algunos autores lo consideran el más grande y completo directorio de la red, mantiene un crecimiento muy rápido. Fue creado como un servicio público y como banco de pruebas para el *Deep Query Manager* (*DQM*), que es un servicio para abonados y una poderosa herramienta para descubrir y gestionar el contenido de internet (Web Profundo y Web Superficial) de forma sin precedentes, flexible y potente.

CompletePlanet cuenta con el más completo listado disponible de todas las máquinas del Web Superficial y de las bases de datos del Web Profundo. Como características principales se pueden señalar además:

- Tiene 103,000 sitios para buscar, organizados en más de 4,000 temas.
- Permite buscar en su directorio o hacer una búsqueda combinando distintas temáticas.
- La estrategia de búsqueda puede ser una lista de términos, una frase o una pregunta escrita en lenguaje natural.
- Permite utilizar los siguientes operadores: AND (&, +), OR, NOT(-), AND NOT, NEAR, BEFORE, AFTER, “”.
- Se pueden usar los paréntesis para agrupar los operadores.

- No limita el nivel de anidamiento en una estrategia de búsqueda.
- Al mostrar los resultados CompletePlanet da un grupo de indicadores sobre el sitio:

  Relevant: Relevancia para la estrategia de búsqueda.

  Popular: Frecuencia con que el sitio es solicitado.

  New: Indica si el sitio ha sido recientemente incorporado.

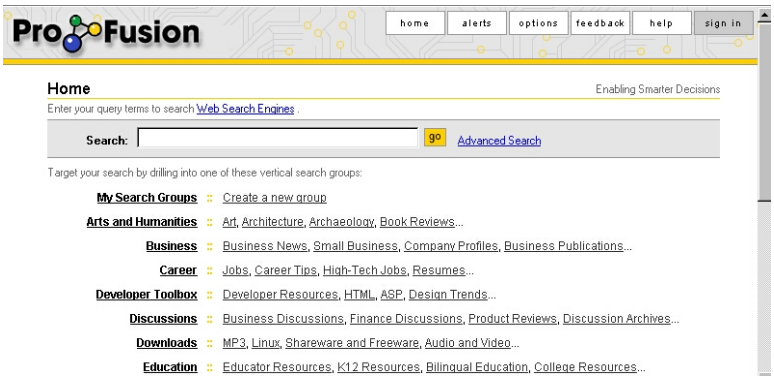
  Link: Presentan los enlaces externos desde el sitio recuperado.

  In DQM: Indica si el sitio es controlado por el DQM.

Pantalla principal del sitio de CompletePlanet

Profusion: www.profusion.com

Este sitio fue adquirido por la compañía de búsquedas Intelliseek en abril del 2000. Se apoya formalmente en la Universidad de Kansas. Busca en algunas de las mayores máquinas de búsqueda del Web Superficial y en un gran número de fuentes en el Web Profundo.



Pantalla principal del sitio de Profusion

Otras herramientas de búsqueda para el Web Profundo

www.fossick.com: Contiene más de 3,000 bases de datos especializadas y máquinas de búsquedas, muchas de disciplinas académicas y tópicos populares.

infomine.ucr.edu: Máquina de búsqueda “académica”, con revistas electrónicas, libros, catálogos de bibliotecas en línea, directorios científicos, entre otros.

dir.lycos.com: Lista de bases de datos referenciales en tópicos científicos y populares.

www.thebighub.com: Indice de más de 3,000 temas de bases de datos específicas agrupadas en más de 300 categorías.

www.webdata.com/webdata.htm: Portal de bases de datos especializado en encontrar, categorizar y organizar bases de datos en línea y proveer enlaces de interés.

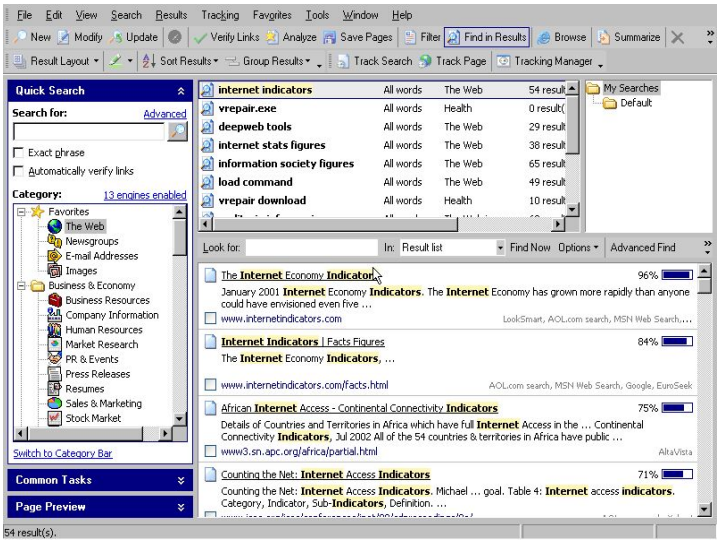
Copernic Agent Pro.
Es un agente inteligente disponible comercialmente, que consulta simultáneamente las más importantes máquinas de búsquedas de Internet. Tiene la versión *Copernic Agent Basic*, que es gratuita.

Copernic AGENT Pro reúne sus búsquedas en más de 120 categorías especializadas, agrupadas entre otras en:

- Favoritos
- Negocios y economía
- Computadoras e internet
- Enciclopedias y referencias
- Gobierno y leyes
- Noticias
- Telemercado (shopping)
- La web

Entre sus principales características podemos citar:

- Facilita la definición de la búsqueda.
- Los documentos son listados de acuerdo a su relevancia
- Resalta en los resultados las palabras buscadas.
- Los resultados duplicados son depurados.
- Las búsquedas pueden ser refinadas usando los operadores: AND, OR, EXCEPT, NEAR sobre los resultados.
- Ofrece una breve descripción de los documentos.
- Las expresiones de búsqueda son almacenadas con los resultados correspondientes.
- Permite diferentes niveles de análisis de los resultados.
- Puede extraer conceptos de las páginas recuperadas.
- Puede consultar más de 1000 máquinas de búsqueda entre las que se destacan: Google, Fast Alltheweb, MSN Web Search, Yahoo, Altavista, Euroseek, AOL.com Search, HotBot, Teoma, Wisenut, Lycos, LookSmart, etc.
- Los reportes de las búsquedas pueden ser generados en formato de páginas web, lo que facilita el filtrado, clasificación y revisión de los documentos.
- Suprime los enlaces “muertos” de los resultados.



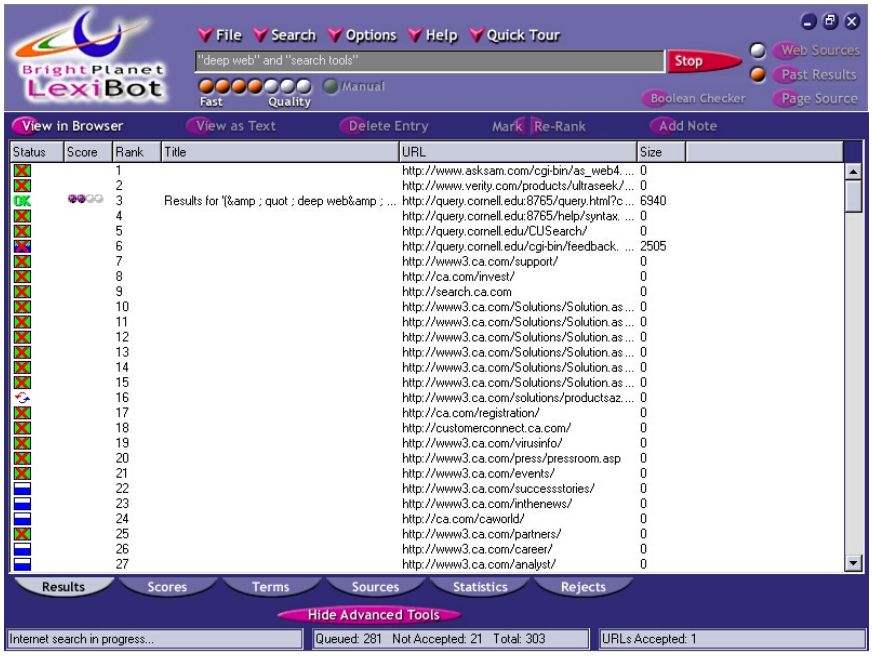
Copernic Agent Pro

LexiBot

Es un agente de búsquedas para el Web Profundo. Es capaz de identificar, recuperar, calificar y organizar contenidos del Web Superficial y del Web Profundo, a partir de una estrategia de búsqueda dada. Actualmente soporta 4,300 fuentes, categorizadas en más de 180 tópicos, permitiendo seleccionar las fuentes donde se va a buscar.

Como otras características se pueden citar:

- Elimina los duplicados y los enlaces muertos.
- Indiza hasta 1000 documentos entre los resultados.
- Clasifica los documentos por orden de relevancia, por omisión.
- Los resultados obtenidos pueden ser agrupados y consultados posteriormente por esta misma herramienta.
- Los resultados pueden ser exportados a un archivo texto con delimitador (coma) o como html. Posteriormente estos archivos pueden ser importados a otras bases de datos o a Microsoft Excel.



Lexibot

Conclusiones

El desarrollo de las herramientas del Web Superficial (con más de 10 años de trabajo) está muy avanzado en sus capacidades de búsqueda y de cobertura de los contenidos.

Las herramientas de búsqueda del Web Profundo (con 3 años de trabajo) tratan de resolver los problemas técnicos que limitan la cobertura y accesibilidad a las fuentes de información que allí se encuentran.

Actualmente no es posible establecer una comparación entre las capacidades de recuperación que ofrecen el Web Superficial y el Web Profundo porque los diferencian, sensiblemente, sus estadíos de desarrollo.

Referencias bibliográficas

1. The Deep Web; Consultada en Enero 2004. Disponible en: http://library.albany.edu/internet/deepweb.html. University at Albany Libraries. Internet tutorials

2. Searching the Deep Web, Consultada en Enero 2004. Disponible en: http://www.dlib.org/dlib/january01/warnick/01warnick.html

3. Deep Web Technologies, Consultada en Enero 2004. Disponible en: http://www.deepwebtech.com/

4. The Invisible Web, Consultada en Enero 2004. Disponible en: http://www.weblens.org/invisible.html

5. How to Choose a Search Engine or Directory; Consultada en Enero 2004. Disponible en: http://library.albany.edu/internet/choose.html

6. Technology white papers: The Deep Web: Surfacing Hidden Value; Consultada en Enero 2004. Disponible en: http://www.brightplanet.com/technology/deepweb.asp

7. Sitio Completeplanet: Consultada en Enero 2004. Disponible en http://www.completeplanet.com/

8. The Invisible or Deep Web: What is Really Out There!; Consultada en Enero 2004. Disponible en: http://library.trinity.wa.edu.au/library/invis/default.htm

9. Searching the Deep Web. Directed Query Engine Applications at the Department of Energy; Consultada en Enero 2004. Disponible en: http://www.dlib.org/dlib/january01/warnick/01warnick.html.