COMITÉ DE COOPERACIÓN BIBLIOTECARIA DE EL SALVADOR

http://www.ues.edu.sv/ccbes/

EVOLUCION DE LOS SISTEMAS DE BUSQUEDA EN INTERNET: LOS PRINCIPALES DIRECTORIOS y MOTORES DE BUSQUEDA

Por: Carlos R. Colindres
Email: : cr.colindres@gmail.com

Resumen

En este artículo, se presentan algunas de las principales formas en que podemos acceder a la enorme cantidad de información disponible en la Red. En particular, se analiza el uso de los *Directorios Web* y de los llamados *Motores de búsqueda* automatizada, presentando un breve resumen de los sistemas más populares hoy en día.

The 'Net is a waste of time, and that's exactly what's right about it.

William Gibson, US science fiction novelist in Canada (1948 -)

The ultimate search engine would basically understand everything in the world, and it would always give you the right thing. And we're a long, long ways from that.

Larry Page, Co-fundador de Google

People Keep asking me what I think of it now that it's done. Hence my protest: The Web is not done!

Tim Berners-Lee, Creador del World Wide Web

Introducción

Una de las principales quejas que se han escuchado siempre acerca de la Internet es la dificultad de encontrar la información específica que se desea. Aunque exageradas, estas quejas no dejan de tener algo de cierto, sobretodo durante los primeros años de la Red, cuando las búsquedas de información podían complicarse enormemente con el uso

de los primeros sistemas[1], que permitían la interconexión de computadoras y la descarga de documentos y archivos.
Sin embargo, en los últimos diez años, ha habido notables mejorías en la forma en que la información es organizada en el *ciberespacio*. La Red se ha convertido en herramienta indispensable para los investigadores y solo se requieren de algunos conocimientos básicos en el uso de los *browsers*[2] y sistemas de búsqueda para aprovechar al máximo los recursos disponibles.

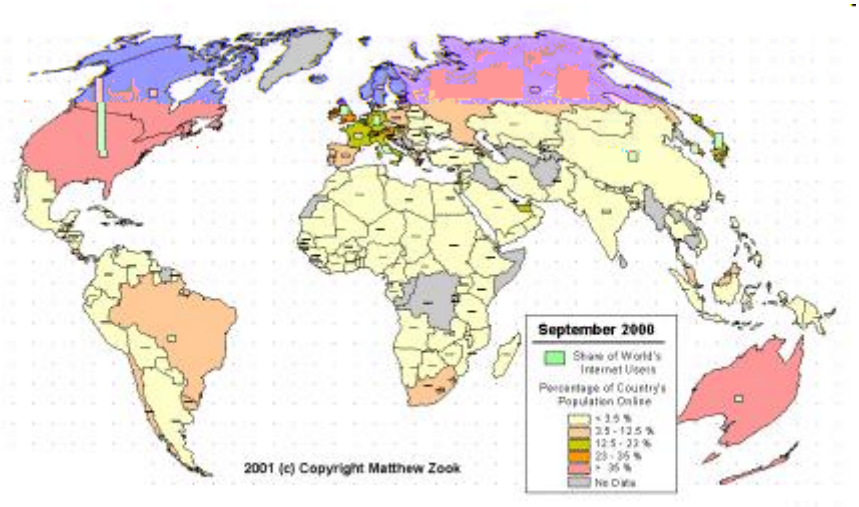

FIG.1 - Distribución en el uso de Internet a Inicios del siglo XXI

En la actualidad, podemos acceder a la información en Internet de diversas maneras siendo las principales mediante el uso de Directorios o Motores de Búsqueda.

DIRECTORIOS

Los directorios son listas de sitios seleccionados por grupos de personas o instituciones en base a algún criterio o temática específica.

Entre los más destacados directorios generales en la Red, se encuentran:

- **WWW Virtual Library**: uno de los directorios más antiguos en la Web, iniciado por Tim Berners-Lee[3] en 1991. Este recurso no- comercial es mantenido

[1] Antes de que se popularizara la World Wide Web (la Web, en corto) a mediados de la década de los años Noventas, los usuarios de Internet contaban básicamente con tres herramientas para intercambiar información: *E-mail* (para el envío personalizado de mensajes electrónicos), *Telnet* (para conectarse a una computadora remota) y *FTP* (para depositar o recuperar archivos de texto o binarios). En el mejor de los casos, el "cibernauta" podía tener acceso a listados de los distintos archivos y carpetas en un servidor en particular conectado a la Red.

[2] Un *browser* es un programa de software que sirve para cargar y desplegar páginas web. El *browser* interpreta código HTML y XML en las diferentes páginas, ejecuta scripts ocultos y programas, provee seguridad a través de encriptamiento/ desencriptamiento. Los principales browsers son Internet Explorer, Netscape, Firefox, Opera, Mosaic y Lynx (este último solo para manipular texto).

[3] En 1989, el británico Tim Berners – Lee propuso a los directores de CERN (Laboratorio Europeo de Física Nuclear) la creación de un sistema de información que facilitara no solamente el acceso sino la publicación de información en la red. Berners-Lee y sus colaboradores desarrollarían en los siguientes

por instituciones voluntarias que se apegan a ciertas normas para incorporar nuevas entradas. La información es ordenada en base a distintas temáticas (por ejemplo, Artes, Educación, Ingeniería, Leyes, Computación, Sociedad, etc) o geográficas. Se estima que tiene en la actualidad un aproximado de 4,000 entradas o enlaces.
URL: http://vlib.org

- **Yahoo!** (Un acrónimo que supuestamente significa *Yet Another Hierarchical Officious Oracle*;): Este directorio de índole comercial fue establecido a finales de 1994 por dos estudiantes de Stanford University. Además de presentar un directorio temático, *Yahoo!* ofrece otros servicios adicionales: cuentas de correo electrónico, anuncios clasificados, venta de artículos, etc. A diferencia del directorio anterior, *Yahoo!* mantiene un motor de búsqueda que permite indicar frases exactas, especificar el tipo de dominio a acceder (.com, .edu, etc), el idioma en que se requiere la información, etc.
  URL: http://www.yahoo.com

Existen además, numerosos directorios especializados, mucho más reducidos y específicos, que pueden tratar sobre casi cualquier temática en particular. A continuación se muestran algunos ejemplos:

- **EServer Journals Collection**: Este directorio incluye acceso al texto completo de más de 400 títulos de publicaciones que cubren un amplio rango de temáticas (cultura, política, informática, literatura, periodismo, comunicaciones, artes, entre otras). Este sitio es mantenido por el Departamento de Comunicaciones Técnicos de University of Washington.
  URL: http://eserver.org/journals/

- **Latin American Network Information Center**: un sitio mantenido por University of Texas y el Instituto para Estudios Latinoamericanos *Teresa Lozano*, dirigido a facilitar el acceso a los múltiples recursos disponibles en la región Latinoamericana. Actualmente, consta de más de 12,000 direcciones únicas ubicados en o que tratan sobre América Latina.
  URL: http://lanic.utexas.edu/

- **WWW en El Salvador**: este Directorio es mantenido por la Universidad de El Salvador, donde se incluyen enlaces a sitios web de instituciones de gobierno, educativas, culturales, no gubernamentales y otras, que tengan dominios salvadoreños[4] o que traten sobre asuntos salvadoreños.
  URL: http://virtual.ues.edu.sv/ref/salvador.html

meses la *World Wide Web*, creando de paso el primer servidor web y lenguaje para diseñar páginas web (HTML- Hyper text Marup Language).

[4] Es decir, que utilicen el código ISO para El Salvador. Por ejemplo: .com.sv; .org.sv; .edu.sv, entre otros.

MOTORES DE BÚSQUEDA

Con la dramática explosión de *hosts* o *servidores* conectados en Red a inicios de los años Noventas[5], varias universidades norteamericanas comenzaron a diseñar nuevas formas de acceder a las enormes y crecientes cantidades de información disponibles en línea. De esta forma surgirían los primeros *motores de búsqueda* (conocidos en inglés como *Search Engines*). Estos sistemas utilizan software conocido como *Web Robots*, *Web Spiders* o *Wanderers* que se encargan de construir automáticamente sus bases de datos partiendo de un rastreo de todos los directorios o páginas web conocidos.

Los motores de búsqueda pueden clasificarse en dos categorías, según su diseño y características generales:

1ª Generación (1990- 1995)

Los primeros buscadores estaban diseñados a partir de herramientas como Telnet, FTP y correo electrónico para implementar la conexión remota y la transferencia de archivos. Entre estos primeros sistemas, se encontraban Archie, Verónica, Gopher y WAIS, descritos a continuación.

- **Archie** (término abreviado de *Archivador*): Creado en 1990, por McGill University de Montreal. Este sistema servía para localizar archivos en servidores FTP públicos, mediante una conexión "anónima". El sistema examinaba periódicamente todos los sitios FTP conocidos, listaba sus archivos y construía un índice. Una vez que el usuario encontraba el archivo que le interesaba, podía copiarlo a su máquina vía FTP, utilizando para ello comandos UNIX para poder desplazarse dentro del sitio. En su mejor momento (alrededor de 1995), las bases de datos de Archie llegaron a indizar más de 2,000,000 de archivos y a tener más de 1000 servidores FTP públicos alrededor del mundo.

- **Gopher**: Fue desarrollado por University of Minnesota en 1991, bautizando el sistema con el nombre de su mascota, el *Golden Gopher*. El sistema Gopher contenía notables mejoras para la recuperación de información por medio de la herramienta FTP. Los servidores "Gopher" organizaban sus archivos en menús, brindando acceso a una interface "amigable" y sencilla de utilizar. El usuario podía digitar el texto a buscar y el sistema respondía con un listado de archivos que probablemente cumplían con el término e búsqueda. Un simple <enter> llevaba al usuario al sitio web donde estaba ubicado el archivo correspondiente. Si el usuario lo deseaba, podía auto-enviarse el archivo mediante Email. Gopher significo una notable mejoría sobre los servidores Archie, llegando a tener hasta 10,000 servidores distribuidos en todo el mundo.

- **VERONICA** (*Very Easy Rodent-Oriented Netwide Index to Computarized Archives*): Fue desarrollado por University of Nevada a partir de los servidores Gopher. Su *Spider* (software de exploración) registraba todos los menús Gopher disponibles en el mundo, recopilando enlaces y agrupándolos en un índice. Llego a ser tan popular entre los investigadores que frecuentemente era imposible acceder a cualquier de los servidores implementados.

[5] Se estima que para 1984, la Internet contaba con cerca de 1000 *hosts* conectados. Para 1992, la cantidad había crecido a más de 1 millón de servidores. Hoy en día, se habla de hasta 400 millones de *hosts* en línea en un momento dado.

- **WAIS** (*Wide Area Information Server*): Este sistema fue desarrollado por la Thinking Machines Corp., también en 1991. WAIS se encargaba de indizar el texto completo de los archivos encontrados durante su exploración de sitios FTP y los almacenaba en una base de datos. Una vez conectado, el sistema le permitía al usuario revisar las entradas en la base de datos, organizada en índices temáticos separados. El sistema llego a tener indizadas más de 600 bases de datos a nivel mundial.

Todos estos sistemas de búsqueda, basados en el uso de comandos de texto, pronto caerían en la obsolescencia al surgir el ambiente gráfico del WWW.

2ª Generación (1995- presente)

Con el desarrollo del ambiente multimedia de la Web a partir de 1991, surge una nueva generación de sistemas de búsqueda en línea, más potentes y más fáciles de utilizar. Estos sistemas utilizan la nueva tecnología del *hiperenlace*[6], que permite enlazar documentos y otro tipo de archivos.

Los nuevos buscadores o motores de búsqueda, utilizan *Spiders* o *Web Robots* más sofisticados que sus antecesores. Dependiendo del tipo de implementación, estos *Webots* toman la información de las siguientes fuentes:

- Título de la página web o las primeras palabras encontradas (por ejemplo, "Bienvenido a mi página")
- Palabras clave (escritas como meta-datos dentro del código HTML de la página web, lo que significa que no aparecen en pantalla)
- Texto completo de cada documento o página web.

Todos los motores de búsqueda en general, se encargan de recopilar la información de los millones de sitios existentes, para luego crear índices que combinan diversos factores como frecuencia de uso de un término en cada página, el idioma del documento, cuantos otros sitios hacen referencia a esta página, la "calidad" y "contenido" del sitio, entre otros criterios.

La gigantesca cantidad de datos recuperada por los *Webots*, es almacenada en las bases de datos de cada sistema, de tal forma que éste pueda responder a una consulta en cuestión de segundos.

Aunque los motores de búsqueda han demostrado ser una herramienta muy poderosa para recuperar información, es importante mencionar algunos aspectos generales que deben considerarse al momento de ejecutar la búsqueda:

! " Los términos de búsqueda deben ser lo más precisos y completos posibles, pues de lo contrario, los resultados arrojados pueden contener cientos de miles de enlaces.

[6] El *hiperenlace* permite colocar vínculo en una página web que nos lleva a otro sitio o recurso también disponible en línea. Los *hiperenlaces* generalmente aparecen como texto sub-rayado e impreso en un color distinto (hipertexto), pero también pueden aparecer como imágenes gráficas o botones. Los *hiperenlaces* pueden enlazar a información en la misma página, en una página distinta, puede activar un archivo de video o sonido, enviar un mensaje de correo electrónico, descargar un archivo, buscar en una base de datos, o enlazar a otro tipo de recursos en Internet.

! " La mayoría de los *buscadores*, no diferencian entre mayúsculas/ minúsculas, ni distinguen acentos diacríticos, diéresis, ni la letra “ñ”.
! " Muchos buscadores no utilizan operadores booleanos (AND, OR, NOT) para construir los términos de búsqueda o si lo hacen, esto ocurre únicamente a nivel de búsquedas avanzadas.

A continuación se analizan los principales motores de búsqueda utilizados en la actualidad:

SISTEMA	BREVE DESCRIPCIÓN
altavista	**Altavista** : Fue uno de los primeros motores de búsqueda en ambiente web, desarrollado en 1995 por Digital Equipment Corporation. *Altavista* fue uno de los primeros sistemas en permitir búsquedas multilingües (particularmente chino, coreano, y japonés), así como en ofrecer un sistema de traducción de texto y reconocimiento de lenguajes (Babelfish) y permitir búsquedas de imágenes, audio, y video. URL: http://www.altavista.com
Google	**Google** : Desarrollado por Larry Page y Sergei Brin, dos estudiantes de Stanford University en 1998, este sistema tiene en la actualidad la más amplia cobertura entre los principales motores de búsqueda (más de 7,000 millones de páginas web revisadas e indizadas, según estimados). *Google* ofrece un sistema avanzado de traducción en varios idiomas, búsquedas por dominio, país, o lenguaje, así como la posibilidad de buscar imágenes. Google no reconoce operadores como AND y al igual que otros sistemas, ignora palabras o caracteres comunes (llamados *términos superfluos*). Tampoco diferencia entre mayúsculas/ minúsculas, ni distingue acentos diacríticos , diéresis n la letra “ñ”. URL: http://www.google.com
Northern Light	**Northern Light** : Es un sistema mantenido y desarrollado por una pequeña empresa llamada Northern Light Corp. desde 1995. Este buscador, al que el usuario debe registrarse previamente, mantiene una colección actualizada de documentos en diferentes áreas particulares (noticias, negocios, economía, entre otros), que incluye artículos de revistas que usualmente eluden los *Webots* de otros sistemas. URL: http://www.northernlight.com
WebCrawler	**MetaCrawler** : Este “meta-buscador” fue puesto en funcionamiento en abril de 1994. Utiliza la tecnología de

"meta-búsquedas" que consiste en buscar automática y simultáneamente la información requerida a través de los principales motores de búsqueda (entre ellos Google, Yahoo!, Looksmart, AskJeeves y otros) arrojando los resultados de cada uno de ellos en forma combinada
URL: http://www.metacrawler.com

MSN Search (Microsoft Network Search) : Es una de las más recientes incorporaciones al mundo de los motores de búsqueda. Forma parte de la estrategia de Microsoft Corp. por penetrar aún más al mercado de Internet. Su motor, llamado *msnbot*, se encarga de recopilar la información de los millones de sitios existentes, para luego crear un índice que combina factores como frecuencia de uso de un término en cada página, el idioma, cuantos otros sitios hacen referencia a esta página, entre otros.
Igual que otros sistemas, permite hacer búsquedas limitando el idioma, dominio o país de origen.

¿Búsquedas Booleanas o PowerSearch?

BÚSQUEDAS BOOLEANAS
Aunque los operadores booleanos[7] se han utilizado para recuperar información de catálogos de bibliotecas y bases de datos desde hace muchos años, no son muy populares entre los usuarios en general por su aparente dificultad.

La mayoría de los Motores de Búsqueda no implementan estos operadores o como se ha dicho, lo hacen solamente para realizar búsquedas avanzadas (ver Cuadro No.1).

Cuadro No.1 – Uso de Operadores Boléanos en la Web

Operador	Utilizado por	Ejemplo
AND ("Y" lógico)	Altavista, Excite, MSN, Lycos, Northern Light	agua AND ecología
OR ("O" lógico)	Altavista, Excite, Google, MSN, Lycos, Northern Light	biología OR bioquímica
NOT (exclusión lógica)	Altavista, Excite, MSN, Lycos, Northern Light	Biología NOT bioquímica
NEAR (proximidad)	Altavista (hasta 10 palabras), Lycos (hasta 25 palabras)	Agua NEAR ecología

[7] El álgebra "booleana" fue desarrollada en el siglo XIX, por el inglés George Boole (1815- 1864). En sus obras "Tha mathematical analisis of Logic" (1847) y "An investigation of the laws of thought on which are founded themathematical theories of logic and probabilities" (1854), Boole proponía resolver argumentos lógicos en un lenguaje que podía ser manipulado y resuelto matemáticamente. Así surgirían las tres operaciones básicas de su "algebra lingüística": AND, OR, NOT, con los cuales llegaría a plantearse el famoso enfoque binario; Sí/ No, Falso/ Verdadero, 1/ 0.

POWERSEARCH

Con la implementación de los Motores de búsqueda, ha surgido una nueva metodología para facilitar la localización de información en Internet. Esta nueva forma de construir expresiones de búsqueda, conocida como PowerSearch, se basa en el uso de operadores matemáticos y otros comandos que permiten especificar no solamente que términos incluir o excluir, sino también otros aspectos como dominio, título, URL, y tipo de enlace (ver Cuadro No.2).

Cuadro No.2 – Uso de PowerSearching en la Web

COMANDOS SENCILLOS		
Operador	**Utilizado por**	**Ejemplo**
+ (incluir términos)	Todos los principales buscadores	agua **+**ecología **+**ecosistema
- (excluir términos)	Todos los principales buscadores	biología **-**bioquímica
“ … “ (incluir frase completa)	Todos los principales buscadores	“teoría de la relatividad”
COMANDOS AVANZADOS		
Operador	**Utilizado por**	**Ejemplo**
title: (búsqueda por título)	Altavista	title: Comisión de Derechos humanos recuperará todos los sitios que tenga en su título principal **Comisión de Derechos Humanos**
intitle: (búsqueda por título)		Intitle: Comisión de Derechos Humanos
host: (búsqueda por tipo de sitio web)	Altavista	host: www.cnn.com recuperará todos las páginas con dominio **www.cnn.com**
site: (búsqueda por tipo de sitio)	Google, Excite	site: www.cnn.com recuperará todos las páginas con dominio **www.cnn.com**
url.host: (búsqueda por tipo de sitio)	Lycos, AlltheWeb	url.host: www.cnn.com recuperará todos las páginas con dominio **www.cnn.com**
Domain: (búsqueda por tipo de sitio web)	Hotbot, Altavista	domain: edu.sv recuperará todos los sitios con dominio **edu.sv**
url: (búsqueda por dirección electrónica)	Altavista, NorthernLight, Excite	url: www.cnn.com
u: (búsqueda por dirección electrónica)	Yahoo	u: www.cnn.com
inurl: (búsqueda por dirección electrónica)	Google	inurl: www.cnn.com

Como puede observarse en el anterior cuadro, uno de los principales problemas en el uso de los distintos operadores de Powersearching, es la falta de estandarización, aunque se espera que en el futuro próximo esto sea resuelto.

Existen otros comandos implementados en forma particular por cada buscador, por lo que se recomienda al usuario investigar las opciones avanzadas de cada uno de ellos.