Acimed 2007; 16(4)

Minería Web: un recurso insoslayable para el profesional de la información*

Lic. Sady C. Fuentes Reyes[1] e Ing. Marina Ruiz Lobaina[2]

Resumen

Se estudian los principales conceptos relacionados con la minería Web (Web mining) y se enfatiza en la minería de uso del Web (Web usage mining). Se muestran, además, los resultados obtenidos con la aplicación de la herramienta *Sawmill V.7.0*, utilizada para el procesamiento de ficheros Log.

Palabras clave: Minería Web, minería de uso Web, software.

Abstract

The main concepts related to Web mining are studied, and emphasis is made on the Web usage mining. The results obtained with the application of the Sawmill V.7.0 tool, which is used for processing Log files, are made known..

Key words: Web Mining, Web Usage Mining, software.

Cita (Vancouver): Fuentes Reyes SC, Ruiz Lobaina M. Minería Web: un recurso insoslayable para el profesional de la información. Acimed 2007;16(4). Disponible en: http://bvs.sld.cu/revistas/aci/vol16_4_07/aci111007.htm [Consultado: día/mes/año].

".... en la antigüedad, el hombre occidental quería ser sabio; luego el hombre moderno quiso ser conocedor; el hombre contemporáneo parece contentarse con estar informado (y posiblemente el hombre futuro no esté interesado en otra cosa que en tener datos)."

Iraset Páez Urdaneta

World Wide Web es un medio de difusión económico y de gran importancia en el entorno empresarial. Ante el acelerado crecimiento del World Wide Web y de la competencia entre las organizaciones ha surgido la necesidad de mejorar la calidad de los sitios Web, esencialmente sobre la base del comportamiento de los usuarios que lo utilizan.

Para el descubrimiento de información útil en el Web, la denominada minería Web es una herramienta útil para el hallazgo de nuevos conocimientos; para eso, emplea la información obtenida de los documentos y servicios Web (textos, imágenes, videos, hiperenlaces, ficheros Log, etc.). A continuación, se realiza una panorámica sobre la minería Web, con énfasis en la minería de uso Web, y finalmente se exponen los resultados obtenidos en el procesamiento de los ficheros Log de un servidor Web, por medio de *Sawmill V.7.0*, una herramienta de software para estos fines.

Minería Web

En el ámbito del acceso, recuperación y organización de información, la minería Web es un campo importante de aplicación en Internet. Se utiliza para el estudio del comportamiento de ciertos aspectos esenciales para mejorar la arquitectura de un sitio ayuda a descubrir conocimientos potencialmente útiles a las organizaciones.

Etzioni define la minería Web como el empleo de las técnicas de la minería de datos —data mining (DM) — para descubrir y extraer información automáticamente del Web.1 Entre sus campos de aplicación principales se encuentran:2

- Los motores de búsqueda.
- El comercio electrónico.
- El diseño Web.
- El posicionamiento Web.
- La seguridad.

La minería Web se subdivide en áreas que abarcan el contenido del sitio, la estructura de navegación y el comportamiento de los usuarios (fig. 1).

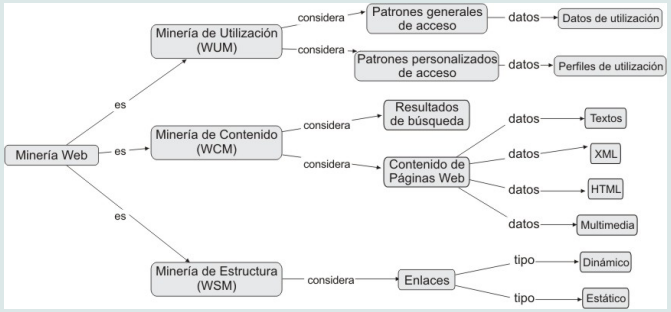


FIG. 1. Mapa conceptual de la clasificación minería Web, según *Juan Carlos Dürsteler*.

Clases de minería

En materia de minería Web, existen tres clases fundamentales:

1. *Minería Web de contenido*. En el Web existen variados documentos, hipertexto, imágenes, vídeos, audio, símbolos, datos, meta-datos, link, textos, pdf y muchos otros, que dificulta su clasificación. La minería de contenido del Web trata de extraer información relevante sobre el contenido del Web, con vista a su clasificación y mejor organización de este, para posteriormente perfeccionar el acceso y la recuperación de la información.
2. *Minería Web de estructura*. Permite conocer cómo se organiza un Web, cómo se estructura y cómo ocurre la navegación en ella.
3. *Minería de uso Web*. Tiene como principal objetivo extraer patrones de uso del Web por parte de los usuarios. Para esto, se utilizan los archivos Log (registros de sucesos/eventos) de los servidores Web. Este tipo de minería tiene dos objetivos principales:
    - Extraer patrones generales de uso de un sitioWeb de manera que pueda reestructurarse para que sea más fácil de utilizar y mejore el acceso por parte de los usuarios.
    - Obtener perfiles de los distintos tipos de usuarios a partir de su comportamiento y navegación, para ofrecer una atención más personalizada.

El procesamiento de Log que se genera automáticamente en los servidores produce información de alto valor. Los datos almacenados en los Log siguen un formato estándar y se almacenan en un archivo de texto, separado cada campo por comas (",") y cada acceso es un renglón distinto.

Entre los datos que registran los llamados Log se encuentran:

- Dirección IP del usuario.
- Fecha y hora de acceso.
- URL de la página accedida.
- Protocolo utilizado para la transmisión de los datos.
- Código de error.
- Número de bytes transmitidos.

Fases

La minería de uso Web presenta cuatro fases fundamentales (fig. 2):

1. *Recolección de datos- búsqueda.* Consiste en la recuperación automática de la información relevante para su posterior procesamiento.
2. *Procesamiento de los datos*. Una vez recuperados los documentos, se ordenan y se preparan para la próxima etapa; se utilizan herramientas para obtener información valiosa en forma automática.
3. *Descubrimiento de patrones.* Existen múltiples técnicas, aplicables al descubrimiento de patrones. Entre ellas, para el agrupamiento y clasificación, para el establecimiento de reglas de asociación y el hallazgo de secuencias frecuentes.
4. Análisis de patrones. Comprende la interpretación y validación de los patrones.

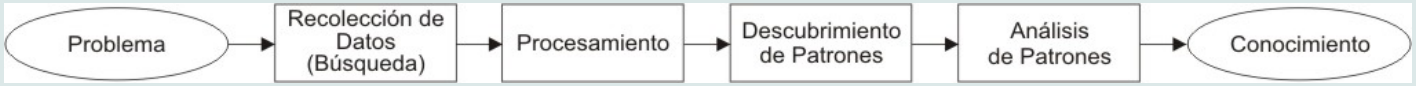


FIG. 2. Fases de la minería de uso Web.

Técnicas empleadas en la minería de uso Web

Entre las técnicas utilizadas se encuentran:[3]

Agrupamiento y clasificación. Las técnicas de agrupamiento o *clustering* distribuyen comportamientos de individuos similares en grupos homogéneos, es decir, dos elementos con características similares pertenecerán al mismo grupo y las características de un grupo (definidas por el elemento prototipo o ideal) serán diferentes a las de otro grupo. En dependencia de la información almacenada en los ficheros Log, es posible detectar grupos de usuarios como:

- Aquellos que visitan gran cantidad de páginas con un tiempo de estancia similar en todas ellas.
- Los que visitan un número pequeño de páginas en sesiones cortas.
- Los que visitan un número pequeño-mediano de páginas con tiempo variable en cada una de ellas.

Una vez descubiertos los prototipos o perfiles de cada grupo, se pueden utilizar las características de cada uno de ellos para realizar la clasificación. En la minería de uso Web, las técnicas de clasificación permiten desarrollar un perfil para clientes/usuarios que acceden a ficheros particulares del servidor, en función de sus patrones de acceso. El agrupamiento de clientes/ usuarios puede facilitar el desarrollo de estrategias de mercado futuras, tanto en línea como fuera de línea. Por ejemplo, envío de correos automáticos a aquellos clientes/usuarios que se encuentren en cierto grupo, reasignación dinámica de servidor para un cliente, tal vez menos sobrecargado, para darle un mejor servicio o la presentación de contenidos específicos según el tipo de cliente.

Reglas de asociación. Las reglas de asociación permiten determinar patrones en los conjuntos de

datos en los que ocurren transacciones de datos. Con esta técnica, pueden encontrarse relaciones sin que exista intervención alguna por parte de algún operador. El descubrimiento de estas reglas ayuda a las organizaciones dedicadas al e-commerce a definir estrategias de mercado efectivas. El aprendizaje de reglas de asociación se divide normalmente en dos fases:

1. Extracción de los conjuntos de ítems que cumplen con la cobertura requerida a partir de los datos.
2. Generación de las reglas a partir de estos documentos.

Secuencias frecuentes. La minería de secuencias permite descubrir el tiempo de las secuencias ordenadas de URLs que han seguido los usuarios y predecir los futuros. En general, en las bases de datos de transacciones están disponibles los datos en un período de tiempo y se dispone de la fecha en que se realizó la transacción. El descubrimiento de patrones de secuencia (*sequential patterns*) en el Log puede utilizarse para predecir las futuras visitas y así poder organizar mejor los accesos y publicidades para determinados períodos de tiempo. Por ejemplo, los días laborables entre las 9 a.m. y las 12 m., muchas de las personas que accedieron al servidor lo hicieron para ver las ofertas, y en los siguientes días la mayoría compró productos. Entonces, por las mañanas se debería facilitar el acceso a las ofertas y brindar la publicidad más llamativa posible.

Herramientas para el análisis de Log

Con el crecimiento explosivo de las fuentes de información disponibles en Internet, es cada vez más necesario que los investigadores utilicen herramientas automatizadas para el hallazgo de los recursos deseados de la información, y así poder conocer y analizar sus patrones de uso.

Para realizar el proceso de extraer conocimiento del contenido de los documentos y de sus descripciones, algo que también se conoce como explotación minera y que permite identificar patrones de comportamiento en los registros de acceso a Internet, existen variadas herramientas. Estas herramientas son sistemas inteligentes que trabajan tanto del lado del servidor, como del lado del cliente, para poder “minar” la información que se genera con el uso de Internet y su análisis se realiza a partir de la información que existe en los archivos Log del servidor de Internet y el servidor de correo. Tenemos entonces:

- *Las que trabajan como herramientas incorporadas al propio servidor*. Estas son aplicaciones del lado del servidor, que corresponden a programas que procesan en tiempo real los datos que se almacenan en los archivos Log. Corren en el servidor, y el acceso a la información del tráfico, tanto estadística, como grafica, se realiza mediante una interfase en línea. Generalmente, este tipo de soluciones vienen incluidas en las ofertas de alojamiento Web, sea un servidor dedicado o compartido.

- *Las que trabajan como herramientas en máquinas personales*. Son software que se instalan de manera independiente en máquinas de escritorio, y su objetivo es igualmente realizar análisis de archivos Log pero no en tiempo real. Esta opción consiste en la descarga de los archivos Log y su posterior procesamiento; por lo tanto, es necesario tener acceso a estos registros, cuestión que debe consultarse con el proveedor del alojamiento Web. Luego, mediante uno de estos programas especializados que se utiliza en una típica PC de escritorio, y sin requerir acceso a Internet, se desarrollan informes estadísticos en poco tiempo. Este es uno de los modos más atractivos y productivos de ejecutar análisis los investigadores del Web.

Cada una de estas herramientas tiene propósitos específicos, como el análisis del uso de la tecnología, el nivel del conocimiento en una institución, las estadísticas de ventas, la usabilidad y muchas otras.

Cabe destacar que cada una de estas herramientas tiene sus propios requerimientos técnicos: espacio disponible en disco, capacidad de memoria, sistema operativo y, por tanto, también

diferentes resultados finales.

Entre las herramientas que trabajan incorporadas al servidor de navegación o de correo, se encuentran: *OmniAnalyzer*, *AWStats*, *Deep Log Analizer V 3.1*, *Advanced Log Analizer*, y *WebLog Expert*. Entre los comerciales, están *DB Miner* y *SpeedTracer*; entre los públicos: *STstat* y *Analog*.

Estudio de caso

La aplicación *Sawmill*

Con el objetivo de ilustrar los aspectos tratados, se estudió el comportamiento de la navegación de los trabajadores del Instituto de información Científica y Tecnológica durante dos días.

Tras una amplia búsqueda en Internet, se determinó escoger como herramienta para el análisis de minería de texto de los archivos Log disponibles el *Sawmill7.2.9_x86_win32* (Demo), que puede utilizarse en servidores de navegación con ISA SERVER Proxy, es decir, que puede emplearse en servidores que generan Log con una estructura diferente a los que genera *Internet Information Server*.

Sawmill es una potente herramienta de análisis de Log. Está especialmente diseñado para analizar Log de acceso a servidores Web, pero puede procesar casi cualquier Log. Se ejecuta como un programa CGI en un servidor Web, y publica un intuitivo interfaz gráfico de usuario, que puede utilizarse desde cualquier navegador para configurar y ejecutar *Sawmill* o para ver estadísticas de páginas. Las estadísticas son jerárquicas, atractivas y poseen enlaces que facilitan la navegación. El programa incluye una completa documentación.

Sawmill ofrece una gran cantidad de opciones, incluida una base de datos persistente, el control sobre la apariencia de las páginas de estadísticas y diversas opciones de filtrado sobre el Log. Este software muestra, tras su instalación, una interfase amigable en *Windows Internet Explorer* y presenta, en un cuadro de selección de opciones ubicado a la izquierda, una serie de estadísticas posibles:

- Cantidad de visitas por hora, por día, por mes, etcétera.
- Horas pico y horas de baja audiencia.
- Páginas más visitadas.
- Páginas de entrada y salida más frecuentes del sitio.
- Utilización de buscadores, clasificación de palabras clave empleadas para buscar.

Algunos resultados

Se procesaron dos días de navegación en el mes de enero del presente año (fig. 3). Se presenta una gráfica y una tabla de la cantidad de visitas realizadas. Estos datos permiten medir el nivel de navegación. Puede observarse, además, que después de las 12:00 del día y hasta las 9:00 de la noche existe un mayor uso (fig 4). En la lista de usuarios, ubicada debajo, puede constatarse si esta carga en el horario señalado se corresponde con la descarga de antivirus y actualizaciones de sistemas.

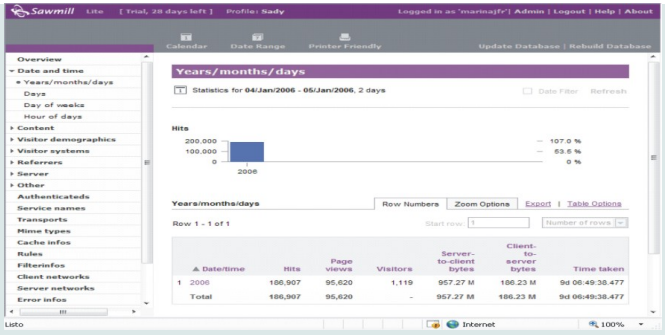


FIG. 3. Informe de uso por año, mes y día.

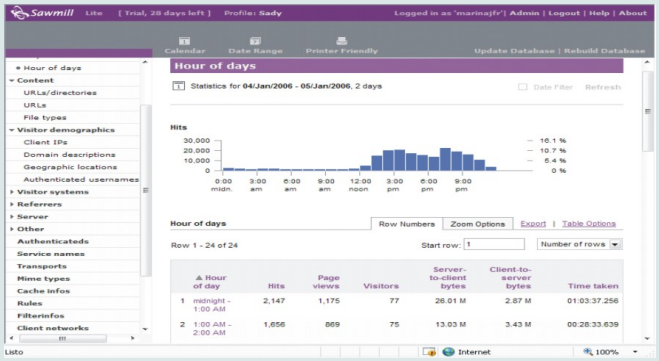


FIG. 4. Estadística de la navegación en el día.

Otra estadística es un desglose por número de IP, la cantidad de entradas, el porcentaje que este representa y el número de páginas visitadas en Internet (fig. 5).

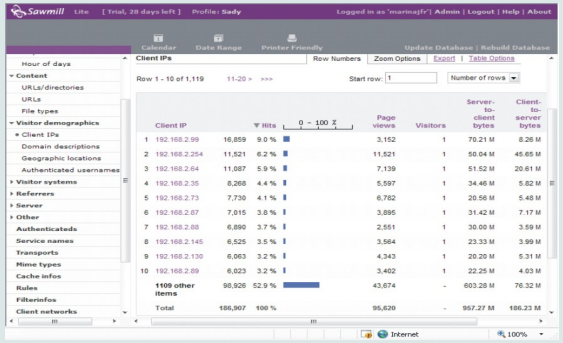


FIG. 5. Informe por cliente IP.

Es posible también hallar las direcciones de Internet visitadas, la cantidad de entradas, el porcentaje que representa y el número de páginas visitadas en Internet (figura 6).

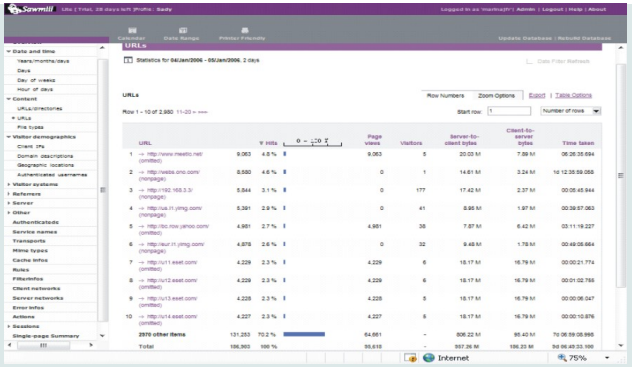
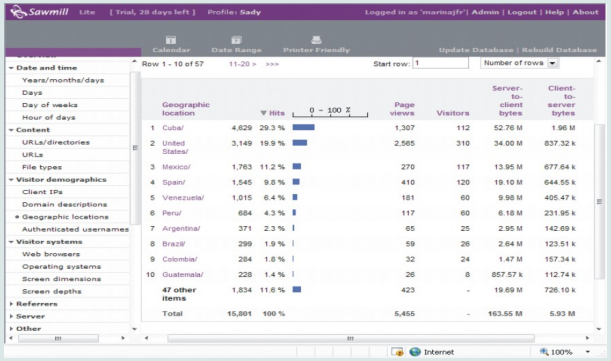


FIG. 6. Informe sobre los URLs.

En el informe denominado *Localización geográfica*, se muestra por país la cantidad de visitas realizadas y el porcentaje que representa (fig. 7). En este caso, el mayor porcentaje de páginas visitadas corresponde a Cuba, seguida de los Estados Unidos.



	Geographic location	Hits		Page views	Visitors	Server-to-client bytes	Client-to-server bytes
1	Cuba/	4,629	29.3 %	1,307	112	52.76 M	1.96 M
2	United States/	3,149	19.9 %	2,565	310	34.00 M	837.32 k
3	Mexico/	1,763	11.2 %	270	117	13.95 M	677.64 k
4	Spain/	1,545	9.8 %	410	120	19.10 M	644.55 k
5	Venezuela/	1,015	6.4 %	181	60	9.98 M	405.47 k
6	Peru/	684	4.3 %	117	60	6.18 M	231.95 k
7	Argentina/	371	2.3 %	65	25	2.95 M	142.69 k
8	Brazil/	299	1.9 %	59	26	2.64 M	123.51 k
9	Colombia/	284	1.8 %	32	24	1.47 M	157.34 k
10	Guatemala/	228	1.4 %	26	8	857.57 k	112.74 k
	47 other items	1,834	11.6 %	423	-	19.69 M	726.10 k
	Total	15,801	100 %	5,455	-	163.55 M	5.93 M

FIG. 7. Localización geográfica.

En otro informe, denominado *Spider* o Araña, se presenta una lista de los motores de búsqueda utilizados y la relación de las páginas visitadas, tanto en forma gráfica como numérica (fig. 8). Se puede apreciar que el motor de búsqueda de mayor demanda es *Google*; *Yahoo*, uno de los motores que lideró Internet, pasa a un tercer lugar.

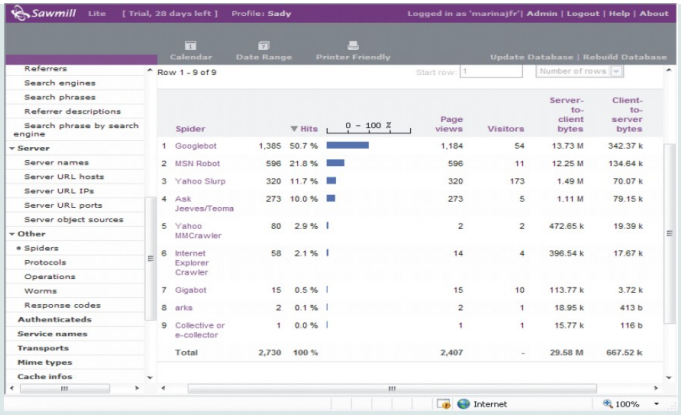


FIG. 8. Motores de búsqueda utilizados.

Con respecto a la clasificación de las palabras más utilizadas, no pudieron obtenerse resultados porque el Log que genera el *ISA Server Proxy* no guarda la frase o palabras clave que se emplearon en las búsquedas. Este puede ser un buen indicador de los temas que se trabajan con más frecuencia en la organización (fig. 9).

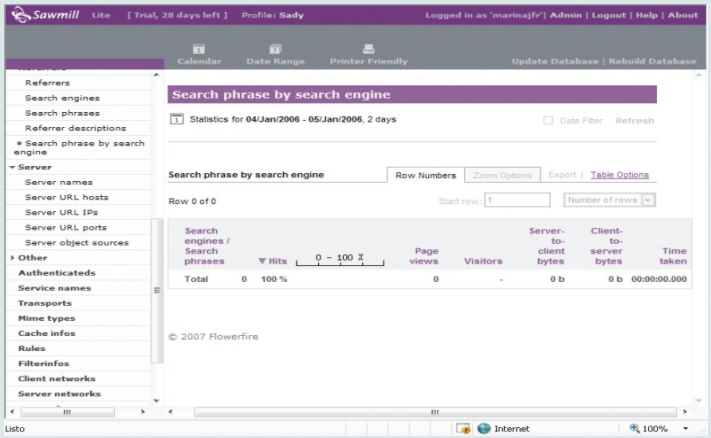


FIG. 9. Frase buscada con motores de búsqueda.

Referencias bibliográficas

1. Ponjuán Dante G. Gestión de información en las organizaciones: principios conceptos y aplicaciones. Santiago de Chile. Centro de Capacitación en Información Prorrectoría. Universidad de Chile. 1998.
2. de Gyves Camacho FM. Web Mining: Fundamentos básicos. Disponible en: **http://zarza.usal.es/~fgarcia/doctorado/iweb/05-07/Trabajos/WMINING.pdf** [Consultado: 22 de agosto de 2007].
3. Montes y Gómez M. Minería de texto: Un nuevo reto computacional. Disponible en: **http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf** [Consultado: 22 de agosto de 2007].

Recibido: 30 de agosto de 2007. Aprobado: 4 de septiembre de 2007.
Lic. *Sady C. Fuentes Reyes*.Instituto de Información Científica y Tecnológica. Instituto de

Información Científica y Tecnológica (IDICT). Capitolio de La Habana. Prado entre Dragones y San José, La Habana Vieja. Ciudad de La Habana, Cuba. Apartado postal 2213. Código postal 10200. Correo electrónico:**cimas@idict.cu**

***Es una edición revisada y ampliada de la ponencia presentada por las autoras en la VI Jornada Bibliotecaria del IDICT, celebrada entre los días 17 y 18 de julio de 2007 en el Capitolio Nacional, La Habana, Cuba. Disponible en: la VI Jornada Bibliotecaria del IDICT, celebrada entre los días 17 y 18 de julio del 2007 en el Capitolio Nacional, La Habana, Cuba.**

[1]Licenciada en Información Científico Técnica y Bibliotecología. Centro de Referencia del Forum de Ciencia y Técnica. Instituto de Información Científica y Tecnológica. [2]Ingeniera Industrial. Departamento Multimedia y Web. Instituto de Información Científica y Tecnológica.

Ficha de procesamiento

Términos sugeridos para la indización

Según DeCS[1]
GERENCIA DE LA INFORMACIÓN; ANÁLISIS DE DATOS; PROGRAMAS DE COMPUTACIÓN; INTERNET.
INFORMATION MANAGEMENT; DATA ANALYSIS; SOFTWARE; INTERNET.

Según DeCI[2]
GESTIÓN DEL CONOCIMIENTO; ANÁLISIS DE DATOS; PROGRAMAS DE COMPUTADORA; INTERNET.
KNOWLEDGE MANAGEMENT; DATA ANALYSIS; SOFTWARE; INTERNET.

[1]BIREME. Descriptores en Ciencias de la Salud (DeCS). Sao Paulo: BIREME, 2004. Disponible en: **http://decs.bvs.br/E/homepagee.htm**

[2]Díaz del Campo S. Propuesta de términos para la indización en Ciencias de la Información. Descriptores en Ciencias de la Información (DeCI). Disponible en: **http://cis.sld.cu/E/tesauro.pdf**

Índice Anterior Siguiente