El web como sistema de información

Lic. Keilyn Rodríguez Perojo[1] **y Lic. Rodrigo Ronda León**[2]

RESUMEN

Se abordan los antecedentes históricos, teóricos y prácticos, necesarios para el surgimiento de una nueva área en las ciencias de la información: la recuperación y la importancia del Web como nuevo espacio para la interacción del hombre con la información hipertextual. Se exponen también, los conceptos Web superficial y Web profundo; se describen algunos de los principales buscadores útiles para explorar el Web profundo, así como las nuevas herramientas para la recuperación de la información como la minería textual y de datos y el descubrimiento de conocimientos en bases de datos.

Palabras clave: Web, procesamiento de la información, recuperación de la información.

ABSTRACT

The practical, theoretical, and historic antecedents necessary for the rise of a new area in information sciences are analyzed: the recovery and the importance of the Web as a new space for the interaction of man with the hypertextual information. The concepts of superficial Web and deep Web are exposed, and some of the main useful search engines to explore the deep Web, as well as the new tools for the information retrieval, such as the textual and data mining and the discovery of know-how in databases are described.

Key words: Web, information processing, information retrieval.

Cita (Vancouver): Rodríguez Perojo K, Ronda León R. El Web como sistema de información. Acimed 2006;14(1). Disponible en: **http://bvs.sld.cu/revistas/aci/vol14_1_06/aci08106.htm** Consultado: día/mes/año.

"Esto que hoy se consideran materiales de biblioteca, las obras del pensamiento y la creación literaria, circularon de forma oral durante mucho tiempo después de la invención de la escritura".[1] El descubrimiento de la biblioteca de *Ebla*, entre las más antiguas que se conoce hasta el momento, revela que las funciones bibliotecarias estaban bien definidas en sus líneas esenciales hace más de 4 500 años:

- Clasificación de los materiales.
- Signaturas en los lomos de las tabletas para su pronta localización.
- Estanterías donde los materiales se ordenan por su forma y contenido para que se conserven con seguridad y se encuentren con rapidez.

El lenguaje documental existe desde la creación de la primera biblioteca, porque este surge cuando el número de volúmenes depositados en un lugar es tan alto que se hace imprescindible su organización de algún modo para permitir la localización de ellos en el momento oportuno y esa organización, desde los orígenes de la biblioteca, se realizó por medio de sistemas rudimentarios de clasificación. El concepto moderno de lenguaje documental debe buscarse a finales del siglo XIX en los aportes de *Melvil Dewey*, autor de la "*Clasificación decimal*" (1876) y *Charles Cutter*, autor del "*Catálogo*

diccionario" (1893), ambos exponentes de dos sistemas documentales:

- La clasificación decimal (*Dewey*).
- La lista de encabezamientos de materia (*Cutter*).

El esquema de la Clasificación Decimal de *Dewe* y, una clasificación decimal que se compone de 10 clases principales, divididas, a su vez, en otras 10, y así sucesivamente, hasta llegar al grado de especificidad considerado deseable, responde a las siguientes características:

- Lenguaje precoordinado.
- Estructura jerárquica.
- Vocabulario controlado.

Esta clasificación, antecedente del resto de las clasificaciones decimales modernas, se ha empleado en bibliotecas de todo el mundo; originó, a su vez, otro sistema aún más popular: la "*Clasificación Decimal Universal*". Asimismo, las teorías de *Cutter* también están vigentes aún, sobre todo las que sirvieron de base a los llamados lenguajes de encabezamientos de materia, caracterizados por ser:

- Lenguajes precoordinados.
- Con estructura asociativa.
- Vocabulario controlado.

Regidos por el principio de especificidad y entrada directa, resultante de su estructura asociativa y alfabética, *Cutter* introdujo una clase de lenguaje documental: los encabezamientos de materia, inéditos hasta entonces y basados en principios completamente diferentes de los que inspiran las clasificaciones:

- El principio de especificidad.
- El principio de entrada directa.

Ambos principios son los pilares en los que se apoya el sistema y rompen con el esquema arbóreo de las clasificaciones bibliográficas; ello, representa un paso de acercamiento al usuario de los sistemas de información. Durante el primer cuarto del siglo XX, se consolidaron las teorías propuestas en las últimas décadas del siglo XIX con la aparición de nuevos sistemas de clasificación bibliográfica y nuevas listas de encabezamientos de materia.

También, comenzaron a aparecer los lenguajes documentales especializados en temáticas particulares, al observarse dificultades en los centros especializados para indizar con los lenguajes enciclopédicos, que no profundizan como es lógico en ninguna temática al intentar abarcarlo todo. Durante este tiempo, el centro principal de la actividad de investigación se desplazó a Europa e incluso a otros continentes. Aparecieron clasificaciones como la de *Henry Evelyn Bliss* (Inglaterra), la de *Brown* (Inglaterra) y la Clasificación Decimal Universal de Paul Otlet (Bélgica). Mención especial merece la clasificación de *Ranganathan* (India), por romper con el esquema de las clasificaciones enumerativas, imperante hasta entonces, y extender el concepto de facetas que tendría repercusión posterior, y que concretamente fue inspiradora del tesauro facetado, aunque realmente fue *Otlet* el primero en ponerlas en práctica.

Tradicionalmente, como resultado del análisis documental, se obtiene una referencia bibliográfica sobre el documento primario que puede constar de los siguientes elementos:

- Una descripción bibliográfica, que incluye datos como el autor, la fecha de publicación, el título, etcétera (noción de metadatos).
- Los términos de indización que representan el contenido del documento y para su posterior recuperación. Estos términos pueden pertenecer al lenguaje libre o natural (no estructurado), proceder de un lenguaje documental, también denominado lenguaje de indización (estructurado y con un vocabulario controlado), o ser una combinación de ambos.

- Códigos de clasificación, que representan temáticamente el contenido por medio de algún esquema de clasificación (también considerado un lenguaje documental).
- Un resumen que representa brevemente el contenido del documento de forma objetiva, es decir, sin interpretación ni crítica. Este resumen puede ayudar al usuario a determinar si el documento realmente es de interés antes de proceder a la consulta del documento primario.

El análisis documental, por tanto, abarca muchas técnicas tradicionales de las bibliotecas, como: la descripción bibliográfica, indización, clasificación y resumen. Combinar las habilidades de los especialistas en información y de los informáticos puede ayudar a organizar el caos existente en Internet. Además, si se considera que el contenido en el Web se encuentra mucho más disperso que en una colección estándar, es comprensible la necesidad de que las habilidades de clasificación y de selección de los sistemas bibliotecarios, se complementen con la automatización de las tareas de indización, clasificación y almacenamiento de la información.

En este sentido, los términos de indización pueden obtenerse por derivación, mediante indización automática, por asignación o por indización intelectual con la utilización de un lenguaje documental externo, por ejemplo un tesauro. En la indización intelectual es un operador humano -generalmente un especialista en información- quien analiza el documento y asigna los descriptores o encabezamientos de materias que considera convenientes.

La asignación de términos de indización consiste en la elección y atribución de términos de indización, aparezcan o no en el texto, para representar documentos o datos, con un lenguaje documental predeterminado. En cambio, en la indización automática es un programa de computadora quien interpreta el documento y asigna los descriptores.

Se trata entonces de una operación compleja en la que intervienen diversas disciplinas como la estadística, la inteligencia artificial, la lingüística, la informática, así como la información y la documentación en función de lograr mejores técnicas para la recuperación de información, porque la exhaustividad y especificidad, o precisión del vocabulario empleado en la indización, influye directamente en la efectividad de la recuperación. La exhaustividad indica en qué grado se registran en el índice del sistema los diferentes aspectos semánticos de un documento, es decir, si los aspectos relacionados con el contenido del documento se registran en el índice con la asignación de un término de indización.

A diferencia de los lenguajes de indización, entre los que ninguno se considera como modelo de referencia, entre los lenguajes de clasificación, existen prestigiosos esquemas, como la *Dewey Decimal Classification* (DDC), la *Universal Decimal Classification* (UDC) y la *Library of Congress Classification* (LCC).

Actualmente, el aumento geométrico del número de documentos existentes en Internet, en particular en el Web y su inestabilidad, ocasionan que los directorios donde la clasificación se realiza por humanos sólo sean capaces de cubrir una pequeña parte del total de recursos existentes en la red y que sean difíciles de actualizar. Es por ello, que los procesos de clasificación automática buscan crear herramientas que ayuden a reducir los costos de la catalogación tradicional mediante la asignación automática de temas a los registros en formato electrónico.

“Con el paso del tiempo, la necesidad de recuperar la información que se encontraba dispersa se hizo evidente, “pero sólo a partir del siglo XX, comenzó a considerarse como un fenómeno de importancia en todos los terrenos” *(Linares Columbié R*. La ciencia de la información y sus matrices teóricas: contribución a su historia. [Tesis para optar por el título de Doctor en Ciencias de la Información]. Universidad de la Habana : Facultad de Comunicación, 2004); así se incrementó el interés por sistemas de indización y clasificación orientados a ambientes automáticos. Con este objetivo, el avance de la ciencia moderna se ha orientado hacia la inteligencia artificial por dos caminos fundamentales: la investigación psicológica y fisiológica de la naturaleza del pensamiento humano, y el desarrollo tecnológico de sistemas informáticos cada vez

más complejos.

La inteligencia artificial, en su sentido más amplio, indicaría la capacidad de un artefacto de realizar los mismos tipos de funciones que caracterizan el pensamiento humano, aplicado a sistemas y programas informáticos capaces de realizar tareas complejas.

LA RECUPERACIÓN DE LA INFORMACIÓN

Acontecimientos tan relevantes para la historia de la ciencia en el siglo XX, como la creación de la primera computadora digital- ENIAC (*Electronic Numerical Integrator and Computer*)- por *John Presper Eckert y John William Mauchly* entre 1943 y 1946, la formulación de la Teoría Matemática de la Comunicación por *Claude E. Shannon* en 1948 y la concepción de una nueva disciplina en el área de la información: la Recuperación por *Calvin Mooers* en 1951, revolucionaron la forma en que se percibía, procesaba, recuperaba y diseminaba la información como activo en las distintas esferas de la economía y la sociedad.

La información, que se trataba desde una perspectiva meramente tradicional (usuario-intermediario-sistema o fondo documental) comenzó a experimentar cambios significativos gracias al desarrollo de las nuevas tecnologías y una *"*amplia proyección de todo tipo de bases de datos en línea*"*.[2]

Desafortunadamente, los distintos modelos de recuperación de información existentes, -un modelo es aquel esquema teórico de un esquema o de una realidad compleja, que se elabora para facilitar su comprensión y el estudio de su comportamiento- conjuntamente con los distintos sistemas de recuperación a los que dieron lugar, no evolucionaron ni mejoraron en la medida que demandaba el crecimiento de la información y la necesidad de acceso a ella. En 1957, tuvo lugar un acontecimiento de innegable repercusión en el ámbito del procesamiento y la recuperación de información: el proyecto *Cranfield I*. Desarrollado en el Cranfield I *nstitute of Technolgy*, constituyó la primera iniciativa para la aproximación a un modelo orientado a crear una metodología para la evaluación de los sistemas de recuperación de información, dicho modelo aún continúa vigente.

Los elementos fundamentales del sistema comenzaron a cambiar la función del intermediario -especialista que participaba activamente en la búsqueda y recuperación de la información- y se delega esta tarea a los sistemas informáticos, sobre la base de que la información procesada por estos, se organiza en forma de documentos hipertextuales, que constituyen *"* las representaciones documentales, entendidas éstas, como un conjunto de caracteres que se agrupan para formar frases y, por último, párrafos y que éstos, de forma más o menos extensa, componen documentos*"*. [3] El usuario, como elemento activo de la consulta y del interrogatorio directo al sistema, representa básicamente su necesidad de información en un proceso en el que se trata de equiparar la representación de los documentos almacenados en la base de datos y los catálogos automatizados con el estado subjetivo de su necesidad de información.

El carácter experimental de estos sistemas de recuperación de información como el " *SMART de Salton* " - un sistema automático de manipulación y recuperación de textos fundamentado en principios estadísticos, cuyo diseño se inició en 1961 por *Gerard Salton* y sus colegas-; el desarrollo de técnicas de retroalimentación como las propuestas por *Rochio* y extendidas por *Amanda Spink*; así como el análisis del concepto de relevancia por *Tefko Saracevic*, constituyen algunos de los ejemplos de que el modelo utilizado y sus técnicas no eran perfectas. En la década de los años 70, cobró fuerza la informática y la recuperación de información se convirtió en un proceso interactivo.

A finales de los años 80 y principios de los 90, comenzaron a materializarse las investigaciones basadas en entornos reales y no en entornos simulados como se hacía hasta aquel entonces. La materialización del Web en los inicios de la última década del pasado siglo por *Timothy Berners-Lee*, el desarrollo del primer navegador web, llamado *Mosaic* en 1993, así como la evolución de las interfaz gráficas de usuario en los

sistemas de búsqueda y recuperación de información, se integran para formar una gran arquitectura de componentes -protocolos, interfaz de aplicaciones, lenguajes de descripción de forma y contenido, etc.-, cuyo funcionamiento dinámico generó lo que hoy se conoce como el Web.

Pero una vez más, la práctica ha demostrado que la interacción entre el usuario del Web y los sistemas de recuperación de información no es efectiva del todo. La débil estructura de la información procesada en bases de datos y sitios Web mediante un esquema de información normalizado, organizado e interoperable que permita una efectiva recuperación de los contenidos, tanto en el llamado Web superficial (*Surface Web*) como en la Web profunda (*Deep Web*), han estimulado a muchos especialistas de distintas áreas del conocimiento -informática, lingüística, inteligencia artificial, psicología, información y documentación, entre otras- a integrar experiencias en favor de crear un Web más organizado. Una de las soluciones posibles es la Web semántica.

Ahora bien ¿por qué la Web semántica? La falta de una infraestructura sólida y estable ha hecho del Web un sistema de información complejo y no muy bien estructurado, donde la gestión, organización, mantenimiento y recuperación de la información se han convertido en un problema para los gestores de información y para el usuario. Como resultado del crecimiento del Web en Internet, se han propuesto distintos mecanismos con el objetivo de reducir las limitaciones de los sistemas de recuperación basados en la navegación hipertextual. Esto ha provocado, a su vez, problemas y limitaciones en los sistemas de recuperación en texto libre, entre ellas:

- Ruido en la recuperación.
- Imposibilidad de acceder a los documentos por campos concretos: autor, temática, fecha, instituciones, etcétera.
- Inadecuación de los métodos de ponderación.
- Sobrecarga del tráfico de la red.

Ante estos problemas, surgió la necesidad de establecer mecanismos para la descripción de recursos, mediante la aplicación de metadatos. El incremento del número de esquemas de metadatos con varios niveles de riqueza y complejidad generados por diferentes comunidades, sean de propósitos específicos o generales, ha ocasionado problemas de interoperabilidad entre estos, porque cada modelo difiere en términos de estructura, sintaxis y semántica.

La interoperabilidad entre metadatos y aplicaciones, definida como la habilidad que poseen dos sistemas y sus componentes para trabajar en conjunto para el intercambio de información de forma eficiente, requiere ante todo del establecimiento de convenciones sobre la semántica, la sintaxis y la estructura de los datos. La semántica se refiere a las necesidades de entendimiento entre esquemas de datos mediante equivalencias del significado mientras que la sintaxis hace referencia a la necesidad de una consistencia sistemática de los datos para el procesamiento por máquina, para el uso y el intercambio de metadatos entre múltiples aplicaciones. Esta última establece restricciones formales sobre la sintaxis para la representación consistente de la semántica. La interoperabilidad de los metadatos y las aplicaciones constituye una de las fortalezas de la Web semántica.

EL WEB COMO SISTEMA DE INFORMACIÓN

La evolución de Internet como red de comunicación global y el surgimiento y desarrollo del Web como servicio imprescindible para compartir información, creó un excelente espacio para la interacción del hombre con la información hipertextual, a la vez que sentó las bases para el desarrollo de una herramienta integradora de los servicios existentes en Internet. Los sitios Web, como expresión de sistemas de información, deben poseer los siguientes componentes:

- Usuarios.
- Mecanismos de entrada y salida de la información.
- Almacenes de datos, información y conocimiento.
- Mecanismos de recuperación de información.

Pudiésemos definir entonces como sistema de información al conjunto de elementos relacionados y ordenados, según ciertas reglas que aporta al sistema objeto- ,es decir, a la organización a la que sirve y que marca sus directrices de funcionamiento- la información necesaria para el cumplimiento de sus fines; para ello, debe recoger, procesar y almacenar datos, procedentes tanto de la organización como de fuentes externas, con el propósito de facilitar su recuperación, elaboración y presentación. Ac tualmente, los sistemas de información se encuentran al alcance de las grandes masas de usuarios por medio de Internet; así se crean las bases de un nuevo modelo, en el que los usuarios interactúan directamente con los sistemas de información para satisfacer sus necesidades de información.

MODELO DE RED COMO MECANISMO DE FLUJO, ORGANIZACIÓN Y RECUPERACIÓN DE INFORMACIÓN EN EL WEB

La nueva manera de entender el espacio urbano está, en opinión de *Gabriel Dupuy* , centrada en el concepto de red "como un conjunto de puntos de transacción, sean estas ciudades, redes técnicas, servicios públicos, redes que generan su propia organización territorial, sin detenerse, en evolución siempre". [4] La red es, no sólo un objeto, sino también una idea globalizadora que expresa la nueva organización del espacio. La idea de red explica mejor que otros enfoques ciertos tipos de relaciones entre el espacio, el tiempo y la información, que se constituyen como elementos esenciales de las sociedades modernas.

Años antes de que *Vannevar Bush* diseñara *Memex y Ted Nelson* acuñara el término *hipertexto* , *Paul Otlet* se refirió a una nueva forma de trabajo en la que por medio de estaciones de trabajo que estarían conectadas en forma de red – *réseau-* , los usuarios podrían compartir información -mediante microfichas en aquella época; así como buscar, leer y escribir a partir de la consulta de grandes bases de datos, cuyo nuevo ámbito de investigación posibilitaría a los usuarios recuperar documentos compartidos en forma de un gran repositorio universal.

La idea de crear redes, tanto desde el punto de vista tecnológico como social, comienza a materializarse con la llegada de Internet a finales de la década de los años 1960, la antesala de lo que hoy se conoce como la red de redes. A principio de la década de los años 1970, un estudiante del Massachussets Institute of Technology (MIT), llamado *Robert Metcalfe* experimentaba con la recién estrenada Arpanet y conectaba computadoras en un laboratorio; con ello, creó lo que llegó a conocerse como *Ethenet*, la tecnología de área local que se utiliza actualmente para conectar a millones de computadoras en todo el mundo. *Metlcalfe*, cofundador de 3Com, hizo la observación de que las redes, bien sean telefónicas, de computadoras o de personas incrementan dramáticamente incrementan su valor con cada nodo adicional. Esto se puede expresar como que la utilidad de una red es equivalente al cuadrado del número de sus usuarios, conocido como Efecto de Red (*Network Effect*).[5]

Internet es un ejemplo válido de la Ley de *Metcalfe*, su rápida expansión en todos los ámbitos de la sociedad así lo demuestra; la red aumenta exponencialmente y, en forma paralela, lo hace su valor. Para *Orihuela*, existen siete paradigmas que caracterizan el nuevo paisaje mediático que emerge en la red:[6]

• Interactividad: La red genera un modelo bilateral, debido a su arquitectura cliente-servidor. Así, los proveedores de contenidos y los usuarios pueden establecer un vínculo bilateral, porque sus funciones resultan intercambiables.

• Personalización: Los servicios de información en línea no sólo se orientan a objetivos con perfiles demográficos, profesionales o económicos similares, sino a individuos, porque la red permite responder a las demandas de información específicas de cada usuario en particular.

• Multimedialidad: La tecnología digital permite la integración de todos los formatos de información (texto, audio, video, gráficos, animaciones) en un mismo soporte.

• Hipertextualidad: Los soportes digitales permiten un modelo de construcción narrativa caracterizado por la distribución de la información en unidades discretas (nodos) y su articulación mediante órdenes de programación (enlaces).

• Actualización: La red posibilita el seguimiento al minuto de la actualidad informativa, y se utiliza en paralelo con la televisión para retransmitir acontecimientos a escala mundial en tiempo real.

• Abundancia: Los medios digitales trastocan el argumento del recurso escaso, porque multiplican los canales disponibles y trasmiten mayor cantidad de información en menor tiempo y a escala universal.

• Mediación: La red cuestiona el paradigma de la mediación profesional de los comunicadores en los procesos de acceso del público a las fuentes y a los propios medios.

Estos paradigmas, que intentan ofrecer una visión razonada de los cambios en los medios de comunicación, potencian nuevos usos y nuevas relaciones en aspectos relacionados con lo económico, lo social y lo cultural. Los nuevos usos se relacionan con la información que fluye por medio de las redes y la manera como ésta se transforma en conocimiento práctico para los usuarios, como es el caso de las redes sociales.

Las redes sociales constituyen un espacio de diálogo y coordinación mediante el cual se vinculan organizaciones sociales e instituciones públicas y privadas en función de un objetivo común y sobre la base de normas y valores compartidos. Estas redes pueden definirse también como un conjunto de personas que representan a organizaciones e instituciones, que establecen relaciones y producen intercambios de manera continua, con el fin de alcanzar metas comunes en forma efectiva y eficiente. Las redes sociales permiten generar relaciones de colaboración, poner en común recursos, desarrollar actividades en beneficio de los participantes, ampliar y estrechar vínculos, crear sentido de pertenencia, socializar conocimientos, experiencias y conocimientos, reconstituir la confianza social, así como establecer relaciones de intercambio y reciprocidad.

Desde el punto de vista tecnológico, uno de los problemas que afronta la visión de intercambio en el Web, radica en la calidad de la recuperación de información una vez más. Buscar información en Internet, con los buscadores tradicionales puede compararse con arrastrar una red en la superficie de un océano: " *no se podrá obtener muchos peces de aguas profundas".*[7]

WEB SUPERFICIAL *VERSUS* WEB PROFUNDO

En 1994, la Dra. *Jill Ellsworth*, especializada en el estudio de Internet, utilizó el término *Web invisible* , por primera vez, para denominar a la información que resultaba "invisible" para los motores de búsqueda convencionales en el Web. También, se denomina "Web profundo" (*Deep Web*), por oposición a la "Web superficial" (*Surface Web)* cuya información puede recuperarse con los buscadores de Internet. La existencia de esta denomina red profunda es un producto de la metodología que utilizan los buscadores para indexar las páginas. El mecanismo se basa en programas llamados robots o arañas, que recorren las páginas de la red siguiendo los enlaces que presentan o se dirigen hacia ellas. Cuando se utiliza alguno de los buscadores conocidos, no se busca en toda la red, sino en su base de datos, construida gracias a la acción de los robots.

A pesar de su pretendida exhaustividad, se calcula que los mayores motores de búsqueda (Google, AlltheWeb) indizan sólo entre un tercio y la mitad de los documentos disponibles para el público en la red. El Web profundo almacena páginas dinámicas que se obtienen como respuesta a interrogantes directas a bases de datos, así como documentos en diversos formatos (mp3. doc, pdf, wma, avi, entre otros), la mayor parte de esta información no se recupera por medio de los directorios y buscadores tradicionales.

En el año 2000, un estudio de la consultora estadounidense BrightPlanet, elaborado por *Michael Bergman* , confirmaba y explicaba la existencia de una red profunda que tendría aproximadamente 7 500 terabytes (equivalente a 7 500 billones de bytes) de información frente a los 19 de la Web superficial o parte de la red accesible mediante los buscadores convencionales.

Actualmente, existen herramientas orientadas específicamente a la labor de recuperar información en el Web profundo como: buscadores, agentes de búsquedas, índices generales y portales verticales. Estas herramientas facilitan el acceso a una mayor parte del Web, porque, además de buscar en el Web superficial, buscan en el Web profundo también, inaccesible para la mayor parte de los buscadores tradicionales. Entre los principales, se encuentran:

- Complete Planet.

Perteneciente a la compañía BrightPlanet, dispone de la lista más completa de todas las máquinas de la Web superficial y de las bases de datos del Web profundo. Creado como un servicio público y como banco de pruebas para el Gestor de Consultas del Web Profundo (*Deep Query Manager o DQM*), que es un servicio para abonados y una poderosa herramienta para gestionar y descubrir contenido en Internet, presenta las siguientes características:

- Posee más de 100 000 sitios para buscar, organizados en 4 000 temas.
- Permite buscar en su directorio o realizar búsquedas mediante la combinación de distintas temáticas.
- La estrategia de búsqueda puede ser una lista de términos, una frase o una pregunta escrita en lenguaje natural.
- Al mostrar los resultados de una búsqueda, CompletePlanet ofrece un grupo de indicadores sobre cada sitio:

- Relevant: relevancia para la estrategia de búsqueda.

- Popular: frecuencia con que el sitio es solicitado.

- New: Indica si el sitio se ha incorporado recientemente.

- Link: Presentan los enlaces externos desde el sitio recuperado.

- In DQM: Indica si el sitio es controlado por el Deep Query Manager (DQM).

- Profusión.

Creado en 1995 en la Universidad de Kansas como un metabuscador inteligente para el Web, fue adquirido por la compañía de búsquedas Intelliseek, en abril de año 2000. Busca en algunas de las mayores máquinas de búsqueda del Web superficial y en un gran número de fuentes en el Web profundo. Permite orientar las búsquedas al definir tópicos generales en los que ellas deben realizarse. También se puede personalizar un grupo de buscadores correspondientes a determinadas materias y obtener resultados de los principales buscadores.

- *Copernic Agent*.

Es un agente inteligente disponible comercialmente que consulta simultáneamente las más importantes máquinas de búsquedas en Internet. Posee la versión *Copernic Agent Basic*, que es gratuita, además de una versión Profesional. *Copernic Agent* Pro, por suscripción y con mayores capacidades de recuperación de la información; reúne sus búsquedas en más de 120 categorías especializadas y entre sus principales características están que:

- Puede consultar más de 1 000 máquinas de búsqueda entre las que se destacan: *Google, MSN Web Search Engine, Yahoo, AOL.com Search* , entre

otras.
- Los informes de las búsquedas pueden generarse en formato de páginas Web, para facilitar el filtraje, la clasificación y la revisión de los documentos.
- Suprime los enlaces muertos de los resultados.
- Puede extraer conceptos de las páginas recuperadas.
- Los documentos se listan según su relevancia.

El desarrollo de las herramientas del Web superficial, cuantitativamente superiores a las herramientas orientadas a la recuperación de información en el Web profundo, las primeras con más de una década de desarrollo y las segundas con alrededor de 5 años de existencia, no pueden resolver problemas técnicos que limitan la cobertura y accesibilidad (en términos de cantidad y calidad) a las fuentes de información disponibles. La sobrecarga de información en el Web supone un gran reto para las organizaciones, especialmente en el manejo de grandes volúmenes de datos para conocer el entorno y predecir su evolución, porque muchas veces poseen la información necesaria para responder a las solicitudes de determinados segmentos de usuarios en el mercado, pero en ocasiones no son capaces de aprovechar al máximo esta información por no tenerla organizada adecuadamente y carecer de los métodos necesarios para procesarla y analizarla de la mejor manera.

PROCESAMIENTO Y RECUPERACIÓN DE INFORMACIÓN: NUEVAS APLICACIONES

Resulta de gran importancia traducir esos grandes volúmenes de datos en información. Desde hace tiempo, es claro que sólo las computadoras pueden manipular rápidamente la inmensa masa de datos y producir informes que apoyen la toma de decisiones. Sin embargo, los resúmenes estadísticos no son la única cosa oculta en el mar de datos. La identificación de patrones comunes, asociaciones, reglas generales y nuevo conocimiento tiene actualmente un gran interés para disciplinas como la minería de texto y el descubrimiento del conocimiento en bases de datos.

Minería textual

La minería textual (*text mining*) es una de las aplicaciones que, desde su formulación a principios de la década de los años 90' del pasado siglo, ha tenido mayor impacto en las actividades de la inteligencia militar. Emplea distintas técnicas de la recuperación de información y la lingüística computacional para facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones documentales. *Marti A. Hearst* en su libro " *Untangling text data mining* ", afirma que ésta tiene como objetivo descubrir información y conocimiento previamente desconocido y que no existe en ningún documento previo.[8]

Relacionada con la minería de datos, la diferencia fundamental radica en que ésta última pretende extraer conocimiento a partir de patrones observables en grandes colecciones de datos estructurados que se almacenan en bases de datos relacionales mientras que la minería de texto realiza la extracción del nuevo conocimiento a partir de grandes volúmenes de información no estructurada. *Hearst* expone que el alcance de la minería textual no está determinado por el desarrollo de la inteligencia artificial propiamente dicha, sino que propone un equilibrio entre el análisis humano y automático a la vez, es decir, un enfoque semiautomático cuyo objetivo intermedio-previo al descubrimiento del conocimiento-es procesar y presentar información disponible en grandes colecciones documentales en un formato que facilite su comprensión y análisis. Entre sus funciones principales se pueden destacar las siguientes:[8]

- Identificar hechos y datos puntuales a partir del texto de los documentos.
- Agrupar documentos similares (*análisis de clusters*).
- Determinar el tema o los temas tratados en el documento mediante la categorización automática de textos.
- Identificar los conceptos tratados en los documentos y crear redes de conceptos.
- Visualización y navegación de colecciones de texto.

La minería textual se presenta como una actividad complementaria a la minería de datos, a pesar de no haber logrado el impacto de esta última. De hecho, existe una similitud entre la minería textual y la de datos, porque ambas persiguen la misma finalidad: deducir información a partir de información existente; cambia sólo el tipo de información que se toma como base de análisis.

Descubrimiento de conocimientos en bases de datos.

El descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Database*) implica un proceso interactivo, que comprende la aplicación de métodos de minería de datos para extraer o identificar aquello que se considera conocimiento, a partir de la especificación de ciertos parámetros en una base de datos. La meta de este proceso es justamente procesar automáticamente grandes cantidades de datos en bruto, identificar los patrones más significativos y presentarlos como conocimiento apropiado para satisfacer las metas del usuario. El proceso de descubrimiento del conocimiento en bases de datos requiere de varios pasos:[9]

- Entender el dominio de aplicación, el conocimiento relevante a utilizar y las metas del usuario.
- Seleccionar el conjunto de datos y enfocar la búsqueda hacia los subconjuntos de variables o muestras de datos donde se realizará el proceso de descubrimiento.
- Filtrar y preprocesar datos, diseñar una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, etcétera.
- Reducir datos y proyecciones para disminuir el número de variables a considerar.
- Seleccionar la tarea de descubrimiento a realizar (clasificación, agrupamiento, regresión).
- Seleccionar el o los algoritmos a utilizar.
- Realizar el proceso de minería de datos.
- Interpretar los resultados.
- Incorporar el conocimiento descubierto al sistema.

Los algoritmos de la minería de datos realizan por lo general tareas de predicción (de datos desconocidos) y descripción de patrones mediante algoritmos de aprendizaje y estadísticos como:[9]

- Análisis de dependencias.
- Identificación de clases (agrupamiento de registros en clases o clustering).
- Descripción de conceptos.
- Detección de desviaciones, casos extremos y anomalías.

Entre los componentes básicos de los métodos de minería de datos están:

- El lenguaje de representación del modelo.
- Evaluación del modelo.
- Método de búsqueda.

La minería de datos ha surgido del análisis potencial de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoyen la toma de decisiones. Por ello, la minería de datos puede clasificarse según las siguientes variantes:[9]

a) Las técnicas aplicadas:

- Sin algoritmos de aprendizaje.
- Consultas SQL (Structured Query Language).
- Procesamiento analítico en línea OLAP (*On- line Transactional Processing*).
- Análisis estadístico (correlación, regresiones).

b) Las funciones que realizan:

- Redes neuronales y algoritmos genéticos.
- Inducción de árboles y reglas.

c) Nuevos algoritmos:

- Inducción de reglas de asociación.
- Inducción de clasificadores bayesianos.

Las diferentes técnicas permiten realizar a sociaciones, clasificaciones, agrupamientos y el establecimiento de patrones secuenciales.

Aunque los diferentes campos de aplicación de la minería de datos demandan el desarrollo de poderosas y costosas herramientas para crear métodos de búsqueda de patrones, no es el único camino existente. El Web, como se conoce hoy, requiere una visión más integral de los problemas de organización y recuperación de información, sobre todo, si se considera que se encuentra estructurado mediante lenguajes de etiquetado que prácticamente describen sólo la forma en que la información debe presentarse al usuario (colores, maquetación, tipografía, etcétera) y dicen muy poco sobre su significado: semántica.

El proyecto denominado Web semántica (Semantic Web) busca que la información pueda reunirse de forma que un buscador pueda comprenderla en lugar de ponerla simplemente en una lista, donde el trabajo que hasta hoy se realizaba en función del usuario (el humano), se centrará en otro tipo de usuario que se valdrá de grandes cúmulos de información, clasificada, descrita y estructurada para una eficiente recuperación: el agente inteligente.

CONSIDERACIONES FINALES

De manera general, se puede afirmar que el Web actual requiere de nuevas formas de organización de la información y el conocimiento para mejorar la capacidad de acceso, uso y recuperación de información. La Web semántica persigue una Web más inteligente, cuyo objetivo es convertir la información en conocimiento sobre la base del marcado semántico y descriptivo no sólo de la información, sino también de los datos, por medio de metadatos, información estructurada y legible automáticamente, sobre la información distribuida en el Web, que proporcionen a las computadoras una mayor capacidad para gestionar y recuperar dichos datos.

REFERENCIAS BIBLIOGRÁFICAS

1. Escolar Sobrino H. Historia de las bibliotecas. Madrid: Pirámide, 1987.
2. Vargas Quesada B, Moya Anegón F de, Olvera Lobo MD. Enfoques en torno al modelo cognitivo para la recuperación de información: análisis crítico. Ciencia da Informaçao 2002;31(2):107-40. Disponible en: **http://scimago.ugr.es/file.php?file=/1/Documents/CInfo-02.pdf** [Consultado: 2 de febrero del 2005].
3. Moya Anegón F de. Los sistemas integrados de gestión bibliotecaria: estructuras de datos y recuperación de información. Madrid: Anabad, 1994.
4. Dupuy G. El urbanismo de las redes: teorías y métodos. Barcelona: Oikos-Tau, 1998. p.35.
5. Downes L, Chunka M. Unleashing the Killer App. Harvard: Harvard Business School Press, 1998.
6. Orihuela JL. Internet: Nuevos paradigmas de la comunicación. Chasqui. Revista Latinoamericana de Comunicación 2002(77). Disponible en: **http://chasqui.comunica.org/** [Consultado: 5 de marzo del 2005].
7. Llanes Vilaragut L, Carro Suárez JR. Para acceder al Web profundo: conceptos y herramientas. En: Congreso Internacional de Información INFO'2004; abril, 12-16; La Habana ; Cuba. La Habana : IDICT, 2004.
8. Hearst MA. Untangling text data mining. Disponible en: **http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html** [Consultado: 9 de marzo del 2005].
9. Morales E. Descubrimiento de conocimiento en bases de datos. Disponible en: **http://dns1.mor.itesm.mx/~emorales/Cursos/KDD03/** [Consultado: 25 de

marzo del 2005].

Recibido: 12 de enero del 2006. Aprobado: 16 de enero del 2006.
Lic. Keilyn Rodríguez Perojo. Red Telemática de Salud en Cuba. Centro Nacional de Información de Ciencias Médicas-Infomed. Calle 27 No. 110 e/ N y M, El Vedado. Plaza de la Revolución. Ciudad de La Habana. Cuba. Correo electrónico: **keylin@infomed.sld.cu**

[1]Licenciado en Bibliotecología y Ciencia de la Información. Red Telemática de Salud en Cuba (Infomed). Centro Nacional de Información de Ciencias Médicas-Infomed.
[2]Licenciado en Bibliotecología y Ciencias de la Información. Facultad de Comunicación. Universidad de La Habana.

Ficha de procesamiento

Términos sugeridos para la indización

Según DeCS[1]

INTERNET; ALMACENAMIENTO Y RECUPERACIÓN DE LA INFORMACIÓN .

INTERNET; INFORMATION STORAGE AND RETRIEVAL .

Según DeCI[2]

INTERNET; WWW; CLASIFICACIÓN; INDIZACIÓN, RECUPERACIÓN DE LA INFORMACIÓN.
INTERNET; WWW; CLASSIFICATION; INDEXING; INFORMATION RETRIEVAL.

[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