IMPORTANCIA Y BARRERAS EN LA LOCALIZACIÓN Y RECUPERACIÓN DE INFORMACIÓN QUÍMICA DIGITAL

Velazquez-Montes, Imelda y Benavides Kuri, Tlacaelel
Facultad de Química. Universidad Nacional Autónoma de México. Universidad Nacional Autónoma de México, Cañaverales 70-33, Rinconada Coapa CP. 14330, México, DF, México.
< ivm@servidor.unam.mx> <http://cosid.fquim.unam.mx>

Palabras clave: Información; Digital; Barreras; Localización; Recuperación.

OBJETIVO

Con el presente trabajo, se presenta la importancia y las barreras de los diferentes métodos de localización y recuperación de la información científica, principalmente en formato digital, a partir de las diferentes fuentes de información existentes a nivel mundial.

MARCO TEÓRICO

Actualmente el acceso a la información a través de diferentes medios como televisión, periódicos, libros, revistas, etc. es algo cotidiano, aunque no siempre es fácil. La búsqueda de información muchas veces se complica, ya que puede implicar el traslado a otros lugares, provocando pérdida de tiempo, altos costos y otras desventajas como obtener información incompleta, insuficiente o, bien, difícil de manejar por la gran cantidad de información disponible.

La información puede obtenerse de diferentes fuentes como impresa o electrónica (Internet, CD-ROM, Multimedios). Cada una de ellas presenta ventajas y desventajas, pero actualmente el uso de fuentes electrónicas es cada vez más común, por los beneficios que ofrece, como facilitar la búsqueda haciéndola más ágil y rápida, además es posible almacenar gran cantidad de información en un pequeño espacio y se puede recuperar fácilmente.

La información digital es fácilmente gestionable por medio de las bases de datos (BDs), con el acceso amigable desde cualquier computadora.

Un sistema de bases de datos es un conjunto de elementos informáticos y humanos que almacenan, mantienen y proporcionan acceso a una información determinada. Las bases están constituidas por cuatro componentes básicos: hardware, software, usuarios y datos (Zacarías, 2002).

La oferta de (BDs) existentes, es muy variada y a la vez diversificada. Existen (BDs) en todas las disciplinas científicas y ramas más importantes del conocimiento aunque existen áreas que disponen de mucha más y mejor información que otras (Martínez, 2001).

Las (BDs) pueden ofrecerse al público en general usando interfases sencillas. Así es posible atender las necesidades de abogados, bibliotecarios, físicos, científicos, investigadores, académicos y otros trabajadores de la información, de cualquier otra disciplina.

Las (BDs) facilitan:
• El almacenamiento de grandes cantidades de información.
• La recuperación rápida de la información
• La organización y reorganización de la información.
• La impresión y distribución de información en varias formas (Zacarías, 2002).

A grandes rasgos, se puede establecer la siguiente clasificación para las BDs:

Referenciales o bibliográficas
Son aquellas BDs que proporcionan las referencias del artículo original y no disponen de la información final, sino que remiten al usuario a otra fuente de información original (un documento u organización) que les permitirá complementar su consulta.

Fuente o texto completo
Este tipo de BDs proporciona la información final, el dato original o el texto completo de la información. No disponen únicamente de referencias a otros documentos, sino que sus registros incorporan los documentos completos. De esta forma se responde directamente a la consulta formulada por el usuario sin necesidad de dirigirlo a otras fuentes de información, por su tamaño son pocas (Martínez, 2001).

El desarrollo de las Nuevas Tecnologías de Información y Comunicación (NTICs), ha favorecido la difusión y recuperación de la información, sin embargo, esto no significa que la búsqueda de la información sea sencilla. La búsqueda de la información implica el uso de descriptores y buscadores, por lo que se requiere conocer las necesidades del usuario, así como la eficacia de los sistemas de búsqueda.

Para conocer las deficiencias en la búsqueda de la información se han hecho investigaciones acerca de la optimización de la interfase, la metodología de los sistemas de búsqueda y también se ha tratado de estudiar al usuario, su comportamiento, para observar las deficiencias, eficiencias y efectividad en la consulta.

La información química enfocada al usuario final, generalmente implica altos costos que derivan del costo de la aplicación, entrenamiento, soporte, el envío al usuario y otros costos asociados con servidores, soporte técnico y actualizaciones de las (BDs). Por esto es importante que la relación costo-beneficio sea optimizada, mejorando la búsqueda de información.

DESARROLLO DEL TEMA

La gran cantidad de información que se ha acumulado a lo largo de los últimos años, puede ser obtenida a través de índices de búsqueda, pero estos tienen grandes deficiencias y si además se habla de información química, esto se complica, ya que se requiere identificar la información y conocer el lenguaje químico, de manera que se tengan descriptores, para que todos sean incluidos en la búsqueda. En química, el lenguaje es un importante criterio que se debe de tomar en cuenta para la búsqueda de información (Tounge, 2002).

Los métodos para seleccionar la información de compuestos, incluyen análisis de diferencia-similitud, métodos QSAR a gran escala y esquema para distinguir unas moléculas de otras; todos estos métodos requieren el escalamiento de descriptores moleculares.

La información química, constituye un alto porcentaje del total de la información disponible en la red, esto representa dificultad en la búsqueda e identificación de contenidos químicos, para ello pueden usarse estrategias para crear mecanismos, para la recuperación de información química, como la creación de docu-

mentos químicos bajo un mismo estándar, así como los procedimientos para identificar y expresar esta información, para mostrar los índices en una interfase amigable de búsqueda.

Para estudiar las barreras del usuario, se pueden hacer encuestas, o por análisis objetivos. Esto último significa que se estudian las actividades del usuario y de ahí se obtiene información del comportamiento de búsqueda (Cooke, 2002).

Para identificar las deficiencias de búsqueda, se requiere crear programas que permitan el estudio del comportamiento de los usuarios, ya sea de forma individual, por grupo o toda la población. Los datos obtenidos se pueden analizar o manipular posteriormente y también pueden estudiarse usuarios entrenados y no entrenados y ver los efectos a corto y largo plazos (Cooke, 2002).

COMO SE BUSCA

Para la consulta de Bases de Datos se deben tener en cuenta los siguientes puntos:

1. Tema a buscar.
2. Elegir las bases de datos adecuadas.
3. Escribir correctamente las palabras clave ("keywords") a buscar, en inglés, sin errores ortográficos.
4. Elaborar una estrategia de búsqueda lo más adecuada posible:
   a. Escribir de manera clara qué es lo que se busca, limitar a campos de interés: Fecha, autor, producto, etc.
   b. Desglosar la consulta en conceptos separados. Incluir sinónimos para cada concepto de búsqueda.
   c. Agrupar los sinónimos utilizando el operador OR (Muñoz, 2001).

Los descriptores o palabras clave, constituyen la parte más importante en la estrategia de búsqueda. Un descriptor o palabra clave, es una palabra o grupo de palabras que sirve para controlar un concepto único, de manera que debe eludir mensajes polisémicos y sinonímicos. Además debe aparecer asociado a otros descriptores cuando existan conceptos relacionado, e incluso, establecer organigramas conceptuales jerárquicos (Muñoz, 2001).

Cuando se busca información química, pueden existir muchas palabras clave diferentes para un mismo concepto, por lo que se pueden delimitar estos descriptores de modo que los compuestos puedan ser identificados fácilmente y que los descriptores sean universales.

Tounge realizó este escalamiento utilizando diferentes bases de datos, que incluían dos compuestos diferentes: fármacos y compuestos que no eran fármacos. De estas bases de datos se eliminaron compuestos de acuerdo a criterios previamente establecidos, de modo que los compuestos, sólo pertenecieran a alguna de las dos categorías mencionadas. Los descriptores se generaron, a partir de dos paquetes comerciales calculándose hasta 313 descriptores por molécula.

Se obtuvieron datos acerca de los descriptores utilizados y con ayuda de la estadística, se desplegaron las distribuciones de los descriptores entre los juegos de datos. Así, se obtuvieron descriptores representativos a partir de estas distribuciones. Se observó que en los compuestos tipo fármaco, había una distribución mayor debido a la complejidad de este tipo de compuestos.

También se observó que los descriptores obtenidos son suficientes, para cubrir por completo la base de datos y aunque se vayan añadiendo más compuestos conforme se descubren, los descriptores ya presentes sirven para los nuevos compuestos.

Con los descriptores que se estudiaron, se creó una escala universal y de ahí se "escalaron" aquellos que tuvieron una varianza significativa.

Las ventajas de estos estudios, es que simplifica la comparación entre juegos de datos dispares además, también se simplifica la identificación de descriptores, que no tengan una varianza significativa y que sólo lleven a correlaciones falsas, provocando pérdidas de tiempo en la búsqueda.

Esta escala absoluta, también tiene la ventaja de simplificar las comparaciones entre los juegos de datos, pero sólo si se hace el mismo escalamiento para diferentes bibliotecas.

También, se pueden crear filtros para facilitar la búsqueda. Uno de ellos, puede marcar los compuestos que estén en un extremo de la distribución del descriptor dado y de ahí inferir algunas propiedades relacionadas, con la información que se está buscando.

Otro filtro simple, que puede ayudar a definir un descriptor es tomar entre el 10 y el 90% del percentil, de la distribución del descriptor dado.

Mediante el análisis de diferentes BDs, se pueden crear descriptores usando la estadística y se puede eliminar la necesidad de definir diferentes escalamientos, para cada aplicación y comparando las Bases de Datos, se puede observar una gran homogeneidad en las distribuciones. (Tounge, 2002)

COMPORTAMIENTO DEL USUARIO

Para estudiar al usuario se pueden usar "Parser". Un parser es un programa o parte de un programa, que interpreta las entradas a una computador, por medio del reconocimiento de palabras clave o analizando la estructura de las oraciones. Este tipo de programas, pueden identificar "actividades" hechas por el usuario (día, hora, navegación, identificación de terminación, etc).

Así, se han desarrollado archivos de este tipo (CrossParse), que pueden ser examinados para determinar la naturaleza de la actividad de búsqueda. Este tipo de programas, pueden estudiar al usuario y ya sea de forma individual o grupal y también permite determinar el número de usuarios activos en el sistema (CrossFire). A través del consorcio con universidades, se han obtenido hasta un total de 85 sitios con acceso al sistema CrossFire y 4500 usuarios activos del sistema. (Cooke, 2002).

En el lenguaje químico, uno de los criterios más importante de búsqueda, es la estructura química. Cooke ha aplicado un parser llamado CrossParse, escrito en Visual Basic, para tratar de mostrar cuan efectivo es el uso del CrossFire por parte de los usuarios. De este modo se puede comparar, el comportamiento del usuario no entrenado, con el que sí recibió el entrenamiento y que a su vez hizo o no, uso del sistema CrossFire.

En los resultados obtenidos por Cook, se observó que el 51% de las búsquedas, fueron hechas por usuarios no entrenados. También se observó que la búsqueda, no sólo se lleva a cabo por estructuras, sino también por reacciones. Los resultados demostraron un incremento del 60% del total de las búsquedas y un aumento en el número de funciones, después del entrenamiento. También hubo incremento en el uso de los sitios. Los cuestionarios apoyaron los resultados de CrossParse, se mostró un mayor uso del Crossfire después del entrenamiento. La mayor parte de los que respondieron el cuestionario, sintieron que su búsqueda fue más eficiente, después del curso del entrenamiento, aunque la información química es la más difícil de localizar, por ser tan compleja.

Los principales impedimentos o barreras de los usuarios, en la localización y recuperación de información científica, en formato digital se pueden clasificar de la siguiente forma:

a) PERSONALES, como la edad, costumbres y falta de confiabilidad, ya que las personas adultas en general, no tienen tanta habilidad en el uso de computadoras y en muchos casos, se limitan a localizarla por los métodos tradicionales, porque les son más familiares; mostrando así mucha resistencia para usar las

NTICs, lo cual, se puede ver reflejado en su productividad, con los resultados observados en personas, que sí las utilizan.

b) EL TIPO DE INFORMACIÓN Y PARA QUE SE REQUIERE. Con frecuencia, los usuarios recurren a buscar información digital, sin tener claridad en lo que buscan y para qué la utilizarán; ya que en ocasiones, es necesario iniciar la consulta en una base de datos multidisciplinaria, para identificar los descriptores apropiados y ubicar la precisión de la consulta y así poder ir directamente a lo que se requiere, obteniendo con ello, mejores resultados.

c) LOS DESCRIPTORES Y EL IDIOMA, juegan un papel importante, en la localización eficiente de la información requerida, ya que los creadores de las BDs, solicitan a los autores de los trabajos originales, que seleccionen los descriptores o palabras importantes, mismas que servirán para localizar ese trabajo, por lo tanto es muy importante la selección apropiada de esas palabras, para la obtención de la información requerida.

d) CONOCIMIENTO DE LA BD Y SU ESTRUCTURA, esto permite dirigirse, por el camino correcto para la obtención de la información adecuada, en forma eficiente, en virtud de que se cuenta con gran variedad de BDs, que incluyen mucha información sobre el tema de consulta, pero es posible que el enfoque no sea el requerido, lo cual en lugar de ayudar, puede complicar el trabajo.

CONCLUSIONES

La búsqueda de la información debe ser lo más sencilla posible. Por ello es importante que los descriptores sean universales y que también cubran ampliamente la base de datos. La información debe estar identificada para facilitar su búsqueda.

La búsqueda de información es más eficiente si el usuario ha sido entrenado y su aprendizaje continuamente es reforzado, de tal forma, que se mantenga actualizado sobre la NTICs, así como de las metodologías para el uso eficaz de la información requerida, logrando así romper las barreras en el uso de las mismas, en su propio beneficio, ya que constituyen una poderosa herramienta para el trabajo docente y de investigación.

Actualmente se tienen más y mejores programas, como el SciFinder, producido por CAS, que simplifican la recuperación de información científica digital, para el área química principalmente.

BIBLIOGRAFÍA

COOKE, F., KOPELEV, N., Approaches to Understanding the Searching Behavior of CrossFire Users, *J. Chem. Inf. Comput. Sci.*, 42 (5), 1016-1027, 2002.

MARTÍNEZ, R., GUADALUPE F., *Acceso a la Información Científica Mundial en el Área Químico-Farmacéutica*. Tesis de Licenciatura. Facultad de Química. UNAM. 2001.

MUÑOZ C., HERACLIO, *Manual de uso para las Bases de Datos de Ingeniería Química*. Tesis de Licenciatura. Facultad de Química. UNAM. 2001.

TOUNGE, B.A., PFAHLER, L.B., REYNOLDS,C.H. Chemical Information Based Scaling of Molecular Descriptors: A Universal Chemical Scale fo Library Design and Analysis, *J. Chem. Inf. Comput. Sci.*, 42 (4) 879-884, 2002.

ZACARÍAS LÓPEZ, Arminda, *Las Bases de Datos más importantes en Química de Alimentos*. Tesis de Licenciatura. Facultad de Química. UNAM. 2002.