Recuperación de información en bases de datos en CD-ROM. Problemática actual

JULIA OSCA LLUCH
Instituto de Estudios Documentales e Históricos sobre la Ciencia.
Universidad de Valencia-CSIC.

Se describe el proceso de recuperación de información en bases de datos en CD-ROM y los problemas que se plantean debido a la diversidad de softwares de recuperación existentes en el mercado. Se hace una descripción y categorización de los operadores más utilizados, indicando la posibilidad y modo de empleo concreto de cada uno de ellos.

PALABRAS-CLAVE: Discos ópticos, CD-ROM, recuperación de información, software de recuperación.

Information retrieval from CD-ROM databases: present difficulties.

The information retrieval process from CD-ROM databases is described and difficulties encountered due to the variety of retrieval software on the market. A description and classification of most commonly used operators is made, indicating in particular the possibilities each one offers and how they are to be used.

KEYWORDS: Optical products, CD-ROM, Information retrieval, retrieval software.

INTRODUCCIÓN

Los discos ópticos constituyen la tecnología actual más avanzada para almacenar, gestionar y distribuir grandes volúmenes de información de naturaleza mixta, es decir que combine textos, gráficos, imágenes y/o sonido. De toda la familia de los discos ópticos, los CD-ROM constituyen en la actualidad unos canales de difusión de la información cada vez más extendidos, fenómeno motivado principalmente por tres factores: Por un lado, la instalación y la distribución de las bases de datos en CD-ROM es sencilla, basta con enviar por correo al usuario un solo disco que se monta fácilmente en la unidad lectora. Por otro lado, una de las principales caracteristicas que posee el CD-ROM es el hecho de que la información que contiene es imposible de modificar y borrar, siendo su manejo muy fácil, lo que

contribuye en gran medida a proporcionar seguridad a los usuarios que lo utilizan. Por último, no hay que olvidar otro factor muy importante a la hora de adquirir información, el factor económico. Por fortuna, y cada vez más, el precio de los discos ópticos es cada vez más bajo, y esto permite que cada vez sea mayor el número de bibliotecas y centros de documentación que adquieren bases de datos en este tipo de soporte[1].

El enorme potencial de futuro de las tecnologías ópticas, vendrían, determinadas principalmente por las siguientes características:

- en primer lugar, puesto que se basan en discos extraíbles, no hay límite a la cantidad y diversidad de información que puede gestionar un solo aparato de lectura, y por otro lado, son tan manejables y transportables como el propio papel.
- en segundo lugar, su densidad de grabación es tan alta que, en un solo disco óptico de 18 gramos de peso y de 12 cm. de diámetro, cabe una enciclopedia de varias decenas de miles de páginas incluyendo millares de ilustraciones.
- por otro lado, a diferencia de las tecnologías magnéticas, son inmunes a campos magnéticos y al polvo, y presentan una gran resistencia a golpes o rozaduras.
- y finalmente, como manejan la información en forma digital pueden ser completamente interactivos.

EL CD-ROM

Como señala Lizasoain[2], CD-ROM es un acrónimo cuyo desarrollo en inglés es Compact Disk-Read Only Memory y que puede traducirse como «disco compacto-memoria únicamente de lectura». Desde este punto de vista, un CD-ROM es un disco compacto que físicamente es idéntico a un disco compacto de sonido (CD-DA; Compact Disc-Digital Audio). Se trata, por tanto, de un disco de policarbonato de 120 mm. de diámetro, 1'2 mm. de espesor y un orificio central para el eje de 15 mm. de diámetro.

La tecnología CD-ROM está sustentada desde el inicio por un entramado de estándares provenientes del CD-Audio[3]. Sin embargo, para dar una cobertura estándar en el nivel propio de CD-ROM, se hacía necesario trabajar en ese sentido también

1. RECODER, M. J.; ABADAL, E., CODINA, L. *Información electrónica y nuevas tecnologías.* (1991). Barcelona, Promociones y Publicaciones Universitarias.
2. BORRÁS, J. A.; CANTOS, C. (1993). Edició de discos CD-ROM. *Item,* 12, 123-133.
3. LIZASOAIN, L. (1992). *Bases de datos en CD-ROM.* Madrid, Paraninfo.

desde el principio. Los aspectos físicos se estandarizaron por Phillips y Sony en 1985 a través de un documento denominado «Libro Amarillo» en donde se contemplan todos los detalles de estructura del disco, formato de datos, corrección de errores, etc. En otoño de ese mismo año, se convocó una reunión de todas las empresas, unas 12, relacionadas en el campo CD-ROM en el Lago Tahoe (Nevada, USA) en el Hotel High Sierra. De ahí salió un comité de estandarización que llevó ese mismo apelativo. El grupo High Sierra (HSG) elaboró un informe definitivo en junio de 1986. Los fabricantes de productos CD-ROM adoptaron rápidamente las propuestas del grupo HSG aunque debía pasar todavía por la certificación de organismos internacionales. El «Libro Amarillo» fue traspasado a la ECMA (European Computer Manufactures Association) en 1987, que consolidó un paso más del estándar definitivo. Después de varios retoques en los foros de la NISO, ECMA, ANSI e ISO, se aprobó definitivamente por la ISO (International Standards Organization) en 1988 con el nombre ISO 9660.

EQUIPO INFORMÁTICO NECESARIO

La configuración mínima recomendada para poder leer un CD-ROM es la siguiente:

- ordenador IBM-PC, AT ó cualquier compatible.
- sistema operativo MS-DOS 3.0 ó superior.
- 640 Kb de memoria RAM.
- disco duro.
- lector CD-ROM con la tarjeta controladora correspondiente.
- software para la búsqueda y recuperación de la información.
- y, opcionalmente, una impresora para poder imprimir las búsquedas realizadas.

SOFTWARE DE CONSULTA DOCUMENTAL

Los fabricantes e informáticos dedicados al mundo de los discos ópticos a menudo dicen que crear el hardware resulta sencillo, siendo el software lo que verdaderamente requiere un gran esfuerzo, por lo que podríamos decir que el hecho de que un disco sea bueno se debe en gran parte a su software.

Muchos productos de publicación CD-ROM están relacionados con bases de datos y, por ende, tienen que realizar el acceso a través de lenguajes de interrogación. Existen numerosos softwares de gestión de bases de datos (tanto relacionales como documentales) en el mercado, pero la mayoría están sintonizados para las

condiciones de los archivos comunes en disco magnético (algunos miles de registros, tiempos de acceso de menos de 50 mseg.). La confusión en el usuario es grande pues tiene que realizar, a menudo, el aprendizaje de innumerables lenguajes que hacen manipulaciones similares pero con gramáticas e interfaces visuales muy diferentes. Algunos mayoristas que comercializan varias bases de datos resuelven un tanto este problema, proporcionando un lenguaje propietario común. El caso contrario también se produce al ofrecer varios mayoristas una misma base de datos.

Para comprender un poco la aparición y proliferación de los distintos softwares de recuperación de bases de datos en CD-ROM, hay que considerar distintos elementos:

- en primer lugar, la distinta versatilidad de los softwares depende de la finalidad que tengan estos para la edición del CD-ROM, ya que algunos no están pensados, por ejemplo, para la edición de catálogos de bibliotecas[4].
- por otro lado, el CD-ROM ha impuesto la necesidad de crear softwares especializados para condiciones extremas que precisan unos algoritmos de acceso y unas estrategias de búsqueda mucho más finas que las habituales[3].
- otro problema superpuesto es que la tecnología CD-ROM ha dejado un importante agujero no estandarizado o regulado en los lenguajes de interrogación. Estos, al ser de responsabilidad más próxima a los propios diseñadores de aplicaciones, han permitido la proliferación anárquica de docenas de lenguajes asociados a productos de distintos fabricantes.

La tecnología CD-ROM no define en realidad más que la posibilidad de expandir sensiblemente la capacidad de archivo en línea de un ordenador (microordenador en la mayoría de casos) y por lo tanto, poder utilizar los recursos corrientes del sistema operativo y de los programas de aplicación. Sin embargo, los parámetros combinados de gran volumen de datos y de baja velocidad de acceso pueden imponer ciertos requisitos en el software de acceso para mejorar la eficacia.

LOS LENGUAJES DE INTERROGACIÓN DE LAS BASES DE DATOS EN CD-ROM.

Entendemos como lenguaje de interrogación, el conjunto de medios y recursos de que disponemos para escribir las consultas que realizamos sobre la base de datos. En el caso de las bases de datos en CD-ROM, estos sistemas están pensados para un usuario final no especialista en informática ni en documentación, por este motivo, los programas que permiten recuperar la información están realizados a través de

4. GARCÍA-RAMOS, L. A. (1991). *Discos ópticos. Tecnologías, productos, aplicaciones.* Barcelona, Edic. Técnicas Rede.

menús de fácil aprendizaje. La recuperación de las referencias puede realizarse, al igual que en la teledocumentación clásica, utilizando múltiples criterios de búsqueda como pueden ser: palabras-clave, autor, nombre de la revista, año de publicación, idioma, etc. Tambien se pueden usar operadores.

Los lenguajes de interrogación en CD-ROM se pueden clasificar en 3 grupos [4]:

- Lenguajes desarrollados específicamente para una aplicación CD-ROM.
- Lenguajes de recuperación genéricos pero orientados específicamente para productos CD-ROM.
- Lenguajes desarrollados para disco magnético adaptados para su uso en CD-ROM (es el caso más frecuente).

ESTRATEGIA DE LA BÚSQUEDA DE INFORMACIÓN. (MÉTODOS Y TÉCNICAS DE INTERROGACIÓN).

Además de conocer la estructura y los campos para interrogar a las bases de datos, es necesario conocer los recursos del lenguaje de interrogación que nos ayudará a formular la estrategia de la búsqueda, esto es, la combinación de instrucciones que elegimos para recuperar la información.

Como más adelante veremos, la mayoría de los programas permiten el empleo de recursos tales como ponderaciones, truncar los descriptores o enmascarar caracteres (por ejemplo, psicol* afecta a cualquier término que comience por dicha raíz, *logía se refiere a cualquier término que finalice en ...logía, etc.). Junto a esto, también la mayoría de los programas permiten efectuar búsquedas en texto libre, es decir, sobre toda la información contenida en los ficheros directos. De esta manera se podrían seleccionar todos los documentos en cuyo resumen (en caso de que la base de datos consultada tenga) aparezca una determinada palabra o frase.

Así pues podríamos definir la estrategia de búsqueda como la serie de acciones o procedimientos encaminadas a obtener un conjunto de referencias adecuadas a una petición de información con la mayor rapidez, economía, exhaustividad y pertinencia posibles.

Los procedimientos concretos varían según el formato de cada base y los programas de recuperación que se emplean en cada caso, pero básicamente todos ellos son similares. El procedimiento más simple para llevar a cabo la interrogación al sistema consiste en teclear un término junto con la instrucción adecuada del lenguaje informático que se emplee:

> FIND PSICOLOGIA, o pulsar la tecla de función adecuada a continuación de escribir la palabra PSICOLOGIA, por ejemplo.

En este caso el sistema buscará todas las referencias que contengan este término en cualquier campo o lugar de los registros correspondientes. Por el contrario, si se especifica el nombre del campo donde se desea que se realice la búsqueda, el sistema seleccionará solo aquellas referencias que contengan ese término en ese campo concreto. A priori, parece lógico pensar que este segundo procedimiento arrojará como resultado el obtener respuestas más precisas. (Ej.: FIND PSICOLOGIA IN DESCRIPTORS, o bien, EN DESCRIPTORES PSICOLOGIA).

Sin embargo, no hay que olvidar que campos como el de descriptores han sido diseñados para ser empleados prioritariamente en la recuperación de la información, y los términos que en ellos aparecen deben (si el proceso de indización ha sido correcto) reflejar el contenido conceptual del documento. Como lo más habitual en las grandes bases de datos bibliográficas es emplear descriptores como términos de indización, los distintos sistemas de recuperación nos ofrecen la posibilidad de elegir, en el índice o en el thesaurus, en caso de que la base de datos tenga uno disponible, uno o varios términos concretos –descriptores– que son los que mejor se adaptan a nuestro tema de interés. Una vez elegidos estos términos, el ordenador nos permite recopilar toda la información recogida bajo cada uno de ellos, formando conjuntos.

La mayoría de las bases de datos y de los lenguajes de recuperación permiten efectuar consultas más sofisticadas en orden a mejorar la eficacia en la recuperación de la información. Independientemente del tipo de lenguaje documental que se emplee, la interrogación a la base puede efectuarse de un modo más preciso si en vez de realizar este tipo de consultas simples, la interrogación se hace más compleja. En definitiva, si se formula una ecuación de búsqueda, entendiendo por este término cualquier combinación de términos de búsqueda (términos existentes en los campos sobre los que se quiere efectuar la búsqueda, como por ejemplo, los descriptores o palabras-clave, nombres propios de autores o revistas, fechas, lenguas, editoriales, instituciones... etc.) y de operadores.

Las relaciones y condiciones que ligan a los términos se expresan mediante los operadores. Estos son un tipo específico de palabras que sirven para combinar términos de búsqueda y, de esta manera, formular expresiones complejas. Sin embargo, así como en lo relativo a las normas y procedimientos a seguir con respecto a los términos hay que tener una visión clara de los procedimientos y lenguajes documentales adoptados por los productores de cada base de datos, lo tocante a los operadores está, en gran parte, en función de los lenguajes informáticos de recuperación de la información que empleen los distintos sistemas (ver tabla I).

Vamos a hacer una descripción y categorización de los más frecuentemente utilizados en las búsquedas realizadas en bases de datos en CD-ROM. La posibilidad y modo de empleo concreto de cada uno de ellos y, por descontado, los códigos y símbolos específicos a emplear dependen de los diferentes sistemas (ver tabla II).

1) Operadores lógicos o booleanos:

El empleo de los recursos del álgebra booleana constituye sin lugar a dudas uno de los instrumentos más poderosos de cara a mejorar la eficacia de la recuperación de la información. Se trata del tipo de operadores más universalmente empleados, y es posible contar con ellos en todas las bases de datos y en los distintos lenguajes de recuperación. Las principales relaciones lógicas usadas son:

- Intersección (AND, Y, &). Se utiliza cuando se quiere que los documentos recuperados traten simultáneamente de los conceptos expresados en los conjuntos que han de realizar la intersección.
- Unión o suma lógica (OR, O, +). Reune todos los descriptores de una misma área o que describen un mismo concepto.
- Negación o exclusión (NOT, NO, !). Se utiliza cuando se quiere excluir un subconjunto dentro de un conjunto más amplio.

En general, se relacionarán con la lógica OR los términos correspondientes a un mismo concepto, con AND los conceptos que deben estar presentes simultáneamente y con NO aquellos que se desea excluir.

2) Operadores de expansión:

Habitualmente, en esta categoría se engloban aquellos operadores que sirven para incrementar el campo de acción de la búsqueda a más de un término mediante la sustitución de caracteres en el mismo. Cuando no existe seguridad en la forma de escribir una palabra o cuando se quiere una palabra con variantes, pueden utilizarse este tipo de operadores. Deben utilizarse con precaución pues pueden dar lugar a «ruido», recuperando documentos no previstos.

A su vez, suele distinguirse entre operadores de enmascaramiento y operadores de truncamiento. Los primeros sirven para remplazar un sólo carácter del término de búsqueda. En este caso, el carácter «?» cumple la función de «comodín», de forma que vale por cualquier carácter (incluida la secuencia vacía), pero uno sólo. Su empleo está especialmente indicado en las variaciones y formas alternativas de un mismo término o género.

Los operadores de truncamiento se emplean habitualmente con las raíces de los términos. El símbolo más frecuente suele ser el asterisco «*». Este operador sustituye a cualquier secuencia de caracteres que siga a lo que le antecede.

Algunos autores engloban a ambos operadores bajo la denominación de operadores de enmascaramiento, distinguiendo entre el simple («?») y el múltiple («*»).

En el caso de las bases de datos en CD-ROM se suele utilizar como operador de enmascaramiento el «?» y como operadores de truncamiento frecuentemente el «*» aunque algunas bases de datos utilizan tambien el «$».

3) Operadores de intervalo, comparación y subrango:

También denominados operadores relacionales, sirven para construir intervalos numéricos.

Su uso es de empleo frecuente con el campo «fecha» o con otros campos de tipo numérico o temporal.

Los más habituales son: igual (=), mayor que (>), menor que (<), mayor o igual que (>=), menor o igual que (<=), entre (..), distinto (<>).

4) Operadores de proximidad o distancia:

Se intercalan entre dos términos para establecer entre ellos una relación de proximidad que puede ser en el mismo orden y seguidos (adyacencia) o con términos entre ellos (distancia).

Se emplean, sobre todo, en los campos en texto libre o en las bases de datos en texto completo. Sirven para definir la distancia máxima admisible que debe separar a determinados términos. Habitualmente adoptan la forma de un número que indica el máximo número de términos intermedios.

5) Operadores de cualificación:

Bajo este calificativo nos referimos a la posibilidad que brindan la mayoría de los lenguajes de recuperación de especificar el campo en el que debe verificarse la condición expresada mediante el empleo del resto de los operadores. Es decir, sirve para indicar al sistema que los documentos a seleccionar deben tener la referencia (simple o adyacente) en el campo especificado. Esta posibilidad de definir el campo de búsqueda aumenta la precisión de las búsquedas y además disminuye el tiempo de respuesta.

Este operador suele representarse mediante la particula EN, IN o su forma algebraica (^, =).

6) Operadores de encadenamiento (referencia a otras selecciones, combinar búsquedas):

La mayoría de los lenguajes de recuperación permiten emplear como elementos de una ecuación de búsqueda los resultados de una o unas precedentes. De esta manera se pueden ir efectuando búsquedas simples y, en función de los resultados de las mismas, combinarlas para obtener una respuesta satisfactoria.

Este operador permite hacer consultas complejas dividiéndolas en consultas más pequeñas. Para cada una de los consultas parciales el sistema nos ofrece el total de registros afectados, y en función de esta información, al final podríamos realizar la búsqueda conjunta.

En este caso la representación del operador utilizado para encadenar las búsquedas ya efectuadas, suele ser el número de la línea que se va a utilizar propiamente (n) o anteponiendo un signo algebraico o letra al número de esta línea (#, .Ln, CB...).

Como vemos, todos los operadores tienen como finalidad última facilitar al usuario la recuperación de la información, aumentando las cotas de precisión de la misma. En cualquier caso, hay que tener presente dos cuestiones:

- En primer lugar, que son los operadores booleanos los instrumentos más poderosos de cara a realizar una búsqueda selectiva y eficaz. Su uso es el más extendido, independientemente del tipo de base de datos que sea y del lenguaje de indización que ésta emplee.
- El campo de los descriptores o palabras-clave, caso de que exista, debe merecer siempre una atención especial por parte del usuario pues la consulta al mismo va a ser el elemento fundamental a la hora de seleccionar información pertinente. Si los productores de la base lo han incluido, se supone que contiene los términos que definen el contenido conceptual del documento original.

La estrategia vendrá determinada por la experiencia de búsqueda del documentalista o el usuario, el conocimiento a fondo de las bases de datos consultadas y la correcta aplicación del programa de recuperación.

CONCLUSIONES

De todo lo expuesto anteriormente se pueden extraer las siguientes conclusiones:

- La actual situación creada por la necesidad de aprendizaje de innumerables lenguajes de recuperación en bases de datos en CD-ROM crea confusión en el usuario y sería deseable que los distintos fabricantes y distribuidores llegaran a una normalización en cuanto a los softwares de recuperación de información.
- Es necesario la unificación en la utilización de los mismos términos o signos para todos los operadores, ya que de esta forma se facilitarían, en gran medida, los trabajos de recuperación. Por ejemplo, en el caso del truncamiento, sería conveniente utilizar siempre el asterico (*) para la sustitución de una cadena de caracteres, ya que es el signo más conocido internacionalmente, y el interrogante (?) para la sustitución de un solo carácter.
- Las funciones más utilizadas deberían estar asociadas siempre a las mismas teclas de función en todos los sistemas (este caso sólo se da en la actualidad con la tecla de función F1 que en todos los sistemas de recuperación va asociada siempre a las pantallas de ayuda).

Del estudio de los distintos sistemas analizados, se concluye que es aconsejable:

- La utilización de los operadores booleanos siempre en inglés (and, or, not) ya que esta forma es la más frecuente en los distintos sistemas (aunque algunos sistemas permiten indistintamente la utilización de los operadores en inglés, español o mediante un signo algebraico).
- Por el contrario, la utilización de los operadores booleanos mediante la representación de un signo algebraico es la menos recomendable, ya que algunos sistemas utilizan el mismo signo algebraico para realizar funciones contrarias (el signo + es utilizado en algunos sistemas para representar la unión y en otros el mismo signo es utilizado para la intersección).

BIBLIOGRAFÍA

RECODER, M. J.; ABADAL, E., CODINA, L. *Información electrónica y nuevas tecnologías.* (1991). Barcelona, Promociones y Publicaciones Universitarias.

BORRÁS, J. A.; CANTOS, C. (1993). Edició de discos CD-ROM. *Item,* 12, 123-133.

LIZASOAIN, L. (1992). *Bases de datos en CD-ROM.* Madrid, Paraninfo.

GARCÍA-RAMOS, L. A. (1991). *Discos ópticos. Tecnologías, productos, aplicaciones.* Barcelona, Edic. Técnicas Rede.

TABLA I.

RELACIÓN DE OPERACIONES MÁS FRECUENTES Y TECLAS DE FUNCIÓN MÁS UTILIZADAS

OPERACIONES	SPIRS	CDKNOSYS	DATAWARE	CHADWICK	DIALOG	BOWKER	ISI
MÁS FRECUENTES							
AYUDA	F1	F1	F1	F1	F1	F1	F1
ABRIR BASE DE DATOS		ALT+D					
CAMPOS	F3	F1	ZONA TRABAJO	ZONA TRABAJO	D.D. HELP	ZONA TRABAJO	ALT+F
BUSCAR	F2	F6, F7	DIRECTO	ALT+B	E.MENU S.	ALT+S	F3
ÍNDICES	F5	F8, F10	F2	ALT+I	W.P.INDEX	ALT+B	ALT+D
THESAURUS	F9				F2		
VISUALIZACIÓN DE REGISTROS	F4	ALT+V	F3+F5	F10		F10	F4
IMPRIMIR	F6	ALT+L	F3+F5	F5	F8	F5	ALT+P
BORRAR BÚSQUEDA	F7	CTRL+B	F5	F3	F9	F3	ALT+C
SALIR	QUIT	ALT+D	F7	MAY+F1	QUIT	QUIT	F5

TABLA II.

TABLA COMPARATIVA DE OPERADORES

OPERACIONES	SPIRS	CDKNOSYS	DATAWARE	CHADWICK	DIALOG	BOWKER	ISI
INTERSECCIÓN	AND	.AND, .Y, &	Y, AND, +	ET, END	AND	AND	AND
UNIÓN	OR	.OR, .O, +	OU, OR	OR, ,	OR, ","	OR	OR
NEGACIÓN	NOT	.NOT, .NO, !	SIN	SAUF, ANDNOT	NOT	NOT	NOT
DE EXPANSIÓN	*	*, ?	*, ?	$, ?	?, ??	$, ?	*
DE CUALIFICACIÓN	IN, =	.EN		=		=	ALT+F
DE ENCADENAMIENTO	#N	.Ln	F4	CB, CS	N	CS	N
DE SUBRANGO	<,>,>=,=<	..	<,>,>=,<=,<> ,..	<,>,>=,<=,=		=,>,<	