Bases de Datos para Recuperación de Información. Un Olvido Curricular

Rosalía Peña
Facultad de Documentación
Universidad de Alcalá de Henares
28871 Madrid
e-mail: rpr@uah.es

Resumen

El gran volumen de información a que estamos sometidos en todos los entornos de la vida aconseja que los contenidos curriculares de las titulaciones informáticas se hagan eco del creciente interés, tanto comercial como teórico, del desarrollo de bases de datos referenciales y sistemas de recuperación que facilitan la selección de las fuentes verdaderamente relevantes para cada necesidad de información.

Los Sistemas de Recuperación de la Información (SRI) son sistemas gestores de bases de datos, y por tanto su estudio entraría dentro de las asignaturas de bases de datos.

Las distintas necesidades de los entornos documentales respecto a los típicos de gestión condiciona las características de los sistemas gestores de ambas y de los modelos subyacentes. El estudio comparativo del modelo relacional y documental puede proporcionar interesantes reflexiones sobre el papel de la abstracción, modelos conceptuales y formales, eficiencia, balance entre metas contrapuestas e influencia de la estandarización en el desarrollo de nuevas tecnologías, que son cinco de los conceptos que ACM e IEEE consideran básicos para la formación de los alumnos de Informática.

1. Motivación

Para minimizar el efecto de la rápida obsolescencia de los contenidos curriculares, debido a la gran velocidad con que se producen los progresos tecnológicos en nuestra área, las recomendaciones curriculares para los estudios universitarios de Informática propuestas de manera conjunta por ACM e IEEE en 1991 [1], establecen muy claramente la necesidad de centrar las exposiciones, realzando lo que llaman (*recurring concepts)* conceptos recurrentes, considerados como aquellos que:

- Impregnan toda la informática, incluso a otras ciencias e ingenierías
- Persisten a través de la historia, independientemente de la tecnología
- Son un hilo conductor que asocia elementos dispares
- Constituyen, por tanto, las herramientas básicas para el futuro desarrollo de esta ciencia

El grupo de trabajo determina 12 conceptos recurrentes que deben ser sedimentados a través de 11 áreas específicas de interés, entre las que está *Bases de Datos y Recuperación de la Información*.

En las directrices generales de los planes de estudio españoles [2] se recoge este área dentro de *Estructura de datos y de la información* como una materia troncal de 12 créditos, la cual suele traducirse en los planes de estudio (entre otras) en una asignatura troncal, bajo la denominación mas o menos directa de *Bases de Datos*. Muchos planes de estudio españoles incluyen adicionalmente otras asignaturas de ampliación en este entorno, bien optativas, bien obligatorias. Pero la segunda parte del nombre del área propuesta por ACM/IEEE91: *Recuperación de la Información* está muy sesgadamente, por no decir nada, representada en los curricula españoles: no se contempla una asignatura bajo este nombre, y tampoco se incluye su estudio en la asignatura de Bases de Datos.

En la década de los ochenta, el temario de la asignatura de Bases de Datos y los libros de texto clásicamente empleados para su estudio, presentaban los modelos jerárquico, red y relacional como alternativas a la representación de una parte de la información del mundo real en una

base de datos [3,4]. Esta diversidad de modelos suponía una buena oportunidad para resaltar algunos de los conceptos recurrentes mencionados por las recomendaciones ACM/IEEE91, concretamente: abstracción, modelos conceptuales y formales, eficiencia, y metas contrapuestas (tradeoff). Date [3] en el capítulo 3 de su 3ª edición de *Introducción a los sistemas de bases de datos* hace un estudio comparativo, muy enriquecedor, de los tres tipos de bases de datos de gestión, mostrando cómo el modelo de datos adoptado por el sistema gestor condiciona el lenguaje de datos y por tanto la interfaz del usuario con el sistema.

Los modelos jerárquico y red han ido perdiendo relevancia, prácticamente han desaparecido del mercado y han ido pasando poco a poco a ocupar los apéndices o últimos capítulos de los libros de Bases de Datos para Informáticos [5,6,7] o han desaparecido totalmente [8], centrando el estudio en el modelo relacional. Esta misma tendencia han seguido los temarios de la asignatura de Bases de Datos. El capítulo 3 antes mencionado desapareció en las siguientes ediciones del libro de Date, pero *Evolución, sus motivos y consecuencias* es también uno de los conceptos recurrentes y entendemos que es interesante presentar el hilo conductor que ha llevado a las bases de datos desde el modelo jerárquico al red y relacional, y deberíamos mantener su docencia con este enfoque, consumiendo el mínimo tiempo posible. Como la docencia de los modelos clásicos del entorno de Bases de Datos de Gestión no es el problema que aborda esta ponencia, dejamos de lado este apunte, que ha surgido al hilo por mantener una estrecha relación con la idea subyacente.

El tema que nos ocupa es la docencia de las bases de datos dedicadas a la recuperación de información contenida en documentos. El modelo relacional, y los sistemas basados en él, son adecuados para tratar conjuntos de datos muy estructurados (entidades bien definidas, con propiedades concretas y valores unívocos) y proporcionan respuestas deterministas a preguntas muy concretas. Digamos que el modelo relacional trata adecuadamente datos típicos de gestión, pero existen otros entornos donde la estructura de los datos y las expectativas del usuario no encajan con los patrones mencionados, concretamente las bases de datos referenciales (aquellas cuya salida es una referencia precisa a la fuente de información, su alcance y su disponibilidad) y documentales (que proporcionan al usuario el documento o una imagen del mismo) se manejan mejor mediante sistemas de recuperación de la información.

En el siguiente apartado tratamos la pertinencia de presentar a los futuros informáticos el entorno de gestión y recuperación de las bases de datos referenciales. En el apartado 3 presentamos el enfoque que proponemos para su estudio. El apartado 4 ubica el tema dentro del temario de la asignatura y se evalúa el consumo de tiempo. El apartado 5 propone un acercamiento práctico.

2. Pertinencia del estudio de entornos referenciales o documentales

En 1983 Salton [9] afirmaba “Desde 1980 el volumen de publicaciones científicas se dobla cada 50 años. Con el impresionante crecimiento de la ciencia y la tecnología de los últimos años, el crecimiento de los conocimientos disponibles se ha acelerado en gran medida”. En aquellos momentos aún no se había producido el acceso masivo a la red Internet, que ha venido a dejar corta esta afirmación.

El desarrollo de la ciencia y la tecnología y también la vida diaria dependen drásticamente del acceso a la información, pero hoy día, el problema es tanto acceder a la información relevante como descartar la información inútil.

Los sistemas de recuperación de la información se ocupan de la representación, organización, almacenamiento y acceso a las fuentes de información, generando bases de datos referenciales y cada vez con mayor frecuencia incluyen un acercamiento directo al documento o al contenido del mismo, colección que estamos denominando base de datos documental. Empleamos aquí el término documento en un sentido amplio que abarca cualquier fuente de información: libros, artículos, imágenes, sonido, mapas, cine, vídeo, páginas web, documentos firmados digitalmente.

Un sistema gestor relacional responde eficazmente a las preguntas concretas del usuario, pero las necesidades de información en la mayoría de los casos distan mucho de poderse formular con precisión, incluso la estructura de la

información referencial no encaja con las características altamente estructuradas de los entornos relacionales.

Desde 1960 se está desarrollando un cuerpo sólido de conocimientos teóricos y prácticos para proporcionar una respuesta adecuada a esta necesidad. Existen sistemas de recuperación comerciales de altas prestaciones avalados por gran cantidad de publicaciones científicas. Por poner un ejemplo concreto: La búsqueda por materia: *Information retrieval* en el catálogo de www.amazon.com, a mediados de febrero de 2001, arroja 1644 libros relevantes.

Las peculiares características de la búsqueda de información en Internet, la nueva legislación sobre el documento electrónico y firmas digitales, y el crecimiento de las publicaciones convencionales propician el desarrollo de nuevos SRI. Existe una necesidad, un mercado real para el futuro profesional informático que no estamos cubriendo en los actuales planes de estudio.

Si aceptamos la definición de Base de Datos *como una colección de datos interrelacionados que se informatizan con el doble objetivo de estar integrados y poder ser compartidos*, no cabe duda que el catálogo de los fondos de una biblioteca es una base de datos y la información residente en el conjunto de servidores de la red Internet también lo es. Por tanto, su estudio podría estar incluido en esta materia. Pero los temarios de bases de datos y los libros de texto que suelen seguirse en estas asignaturas hoy día se centran en el modelo relacional. Sin embargo, los procedimientos y modelos aplicables en el entorno referencial, la interfaz con el usuario, las características del lenguaje de manipulación de datos son completamente distintas.

La comparación de enfoques distintos siempre resulta enriquecedora. Algunos textos hablan de bases de datos multimedia, orientadas a objeto, o deductivas [6,7,8,10]. Los textos buscan ejemplos que complementen el modelo. Algunas veces estos ejemplos resultan muy rebuscados y alejados de las experiencias previas de los alumnos. Por ejemplo el texto de Ryan [8] presenta alternativas para el almacenamiento interno de datos de un sistema de información geográfica. El estudio resulta muy interesante y está bien enfocado, pero los sistemas de información geográfica son muy específicos y mucho mas lejanos a las experiencias previas del alumno que las bibliotecas o que la búsqueda en Internet.

Korth [7] dedica un apartado de un capítulo (21.8), escasamente cuatro páginas (de untotal de 641), a presentar los SRI.

Los libros que estudian los SRI [9,11,12,13-por ejemplo] no suelen ser referencias en nuestras Escuelas y Facultades, quedando de uso exclusivo para los estudios de Biblioteconomía y Documentación, lo que produce una falta de entendimiento entre los profesionales de estas áreas, para los que las Bases de Datos son cosas totalmente distintas.

Entendemos que el estudio comparado de ambos modelos es muy enriquecedor para los estudiantes de Informática, aportando no sólo el conocimiento de un área con un mercado actual, sino también el fomento de capacidades extraordinariamente relevantes para su ejercicio profesional.

El entorno SRI proporciona ejemplos de estructuras de datos, organizaciones de ficheros e interfaces de usuario muy características y enriquecedoras. Adicionalmente, la estandarización existente en ese área ha propiciado claramente el desarrollo del nuevo concepto de biblioteca, más global y más próxima al usuario y es un buen ejemplo del papel que los estándares juegan en el desarrollo de las nuevas tecnologías (otro de los conceptos recurrentes propuestos).

Proponemos un estudio comparativo del modelo relacional con modelos para SRI al estilo del desarrollado por Date para la presentación de los Jerárquicos, Red y Relacional.

El problema habitual cuando proponemos la introducción de un tema, es que los temarios siempre parecen sobrecargados y falta tiempo para su desarrollo, sin embargo, las ventajas en este caso compensan el esfuerzo. Podemos aprovechar que en el momento actual no es necesario detenerse en el manejo de los modelos anteriores de gestión, solo en su estilo, los condicionantes de su estructura y su interfaz de usuario. Y reutilizar el tiempo que antes se dedicaba a un desarrollo mas profundo de estos modelos ya desfasados.

3. Entorno de gestión versus entorno referencial

	GESTION	DOCUMENTACIÓN
Ejemplos	Nómina, Matrícula del alumnado, gestión de almacén, préstamo libros	Aranzadí, Chemical Abstract, Catálogo de la biblioteca, Información en la www
datos	Muchas entidades, con propiedades muy concretas y múltiples relaciones entre entidades	Una o pocas entidades con débiles relaciones entre ellas. Propiedades de entidades muy variables.
Patrones de uso	Altas, modificaciones, bajas, y consultas, distribuidas mas o menos uniformemente en el tiempo. Las consultas son muy concretas y repetitivas ejemplo: expediente de un alumno, listas de una asignatura. (excepto tendencia a permitir consultas ad-hoc, y reusar la información para fines no previstos)	Elevada proporción de consultas. Prácticamente no existen modificaciones. Bajas, si las hay se producen en procesos de expurgo. Mantenimiento y consultas por distintas usuarios no solapan en el tiempo. Las consultas pueden ser concretas o no, y no suelen ser previsibles no repetitivas
Perfil usuarios	Los usuarios que mantienen los datos son los mismos que los que consultan. Típicamente dedican su horario laboral a trabajar con la base de datos	Dos tipos de usuarios a) los que mantienen los datos, le dedican parte considerable de su jornada. Conocen bien el sistema b) clientes, exclusivamente para consultas esporádicas a la base de datos

TABLA1: Puntos de partida de entornos de gestión y documental

Las características del SGBD, la estructura interna de los datos almacenados y del lenguaje de usuario vienen determinadas por:

1) Por la tipología de los datos que almacena

2) El perfil de los usuarios, su homogeneidad, su formación, horas de dedicación al SGBD.

3) Por los patrones de uso de los datos, es decir la frecuencia y el momento en que se producen las altas, bajas, modificaciones y consultas.

En la tabla 1 se presentan los puntos de partida de ambos entornos respecto a estos tres aspectos. Las diferencias encontradas propician características peculiares en el lenguaje de recuperación, en las estructuras de datos, en la gestión de la base de datos

1) Lenguaje de recuperación:

. El usuario genera la pregunta mediante expresiones del álgebra de Bool

. Mantenimiento de un histórico de consultas para posteriores perfeccionamientos

. Limitadores que permitan reducir el ámbito de una pregunta

. El usuario no necesariamente indica en qué campo de la base de datos desea buscar

. Dispone de operadores de adyacencia y comodines para flexibilizar las consultas

. Búsquedas en texto libre

. Ayudas para la navegación a través de los datos: un visualizador (browsing) e hipertexto

. Establecimiento de un tesauro de términos de catalogación o lenguaje controlado que facilita la navegación así como la ampliación o reducción del ámbito de las consultas

. Definición de perfiles de usuario, área de interés, documentos que interesaron a otros usuarios que accedieron a este documento.

. Distintas interfaces para distintas necesidades del usuario

3)Estructuras de datos

. Manejo de registros de longitud variable (típicamente evitados en entornos de gestión)

. Ficheros invertidos

. Índices multiatributo

4)Gestión de la Base de datos

. El sistema trabaja mayoritariamente sobre los índices, accediendo al fichero de datos solo al finalizar el proceso

. Cumplimentación de la base de datos de forma automática o semiautomática: generación del tesauro y catalogación

. Redefinición y perfeccionamiento de la consulta del usuario.

4. Planificación temporal

Durante 5 años hemos impartido este tema en la única asignatura bases de datos de las titulaciones de Ingeniero Técnico en Informática de Sistemas y de Gestión en la Universidad de Alcalá de Henares de 60 créditos, consumiendo 4 sesiones de clase de una hora.

El desarrollo del tema propuesto responde a:

BASES DE DATOS REFERENCIALES
. El mercado de la información
. Requisitos condicionados por el entorno
. Características del lenguaje de interrogación:
  Lógica booleana,
  Histórico y refinamiento de consultas,
  Comodines, adyacencia, búsqueda en texto libre
  Tesauro
  Navegación e Hipertexto
.Organización física de la información:
  Ficheros invertidos
  Índices multiatributo.

Se ha insertado en el temario cuando el alumno ya trabaja con soltura con el entorno relacional: modelo de datos, requisitos de normalización y lenguajes de definición y manipulación de datos. Y antes de abordar otras consideraciones de calidad e integridad de datos y de abordar la gestión distribuida.

Al ser un tema que no figura en los libros habituales de consulta de nuestros alumnos hemos desarrollado unos apuntes que pueden servir de base para la preparación de la docencia (si son de su interés los puede solicitar a la dirección de correo indicada en este documento).

Caso de disponer de mas tiempo (entre 4 y 6 horas adicionales) resultaría interesante incorporar un segundo tema en que se abordara la inclusión del documento en la base de datos. Los dos primeros años lo incorporamos. En aquel momento los alumnos mostraban gran interés en estudiar la manipulación de una imagen en movimiento, los procesos de digitalización y OCR, sin embargo, la reducción de tiempo disponible para el resto de materias resultaba demasiado drástica.

El desarrollo del tema podría responder al siguiente esquema:

BASES DE DATOS DOCUMENTALES
. Objetivo: Difusión+conservación de originales
. Textos e imágenes en blanco y negro
. Imagen escala de grises y color
. Sonido
. Imagen en movimiento
. SGBD referenciales, documentales y multimedia
. Sistemas mixtos SGBD-SRI

Por supuesto también es posible montar una asignatura de bases de datos adicional a la introductoria que verse íntegramente sobre SRI, pudiendo en este caso estudiar distintos modelos, algoritmos para retroalimentación por relevancia, lenguaje natural, búsqueda en la web, bibliotecas digitales. Los libros de Baeza y Korfhage [12,13] podrían servir como textos básicos.

5. Prácticas

En el plan de estudios de Alcalá de Henares se imparte paralelamente a la de Bases de Datos una asignatura de Laboratorio de 30 créditos estructuradas en sesiones de horas.

Nuevamente el recurso tiempo es escaso, consideramos que la creación de una base de datos en el entorno referencial aporta menos que una aproximación a su consulta, por lo que no es necesario, siquiera la disponibilidad de un sistema gestor en este entorno. Es suficiente acceder a una base de datos suscrita la Universidad o en su defecto incluso al propio catalogo de la biblioteca o cualquier otro centro documental: la Biblioteca Nacional, Amazón o Medline.

A fin de no estar sujetos a los retrasos que Internet puede provocar en una sesión docente optamos por enfrentar a los alumnos con la consulta de una base de datos construida sobre Knosys, concretamente la Base de datos de Jurisprudencia de Aranzadi. Knosys proporciona suficientes de los elementos que hemos comentado como para suponer un acercamiento razonable al lenguaje de recuperación documental.

Usamos una sesión de 2 horas. En la primera de las cuales hicimos un rápido recorrido por diversas opciones con la ayuda de un cañón de proyecciones, en la segunda hora hicieron ellos búsquedas. Entendemos que incluso solo la primera aproximación propuesta es beneficiosa y enriquecedora si se quiere disminuir la dedicación a una hora.

6. Conclusión

Este trabajo justifica la necesidad de acercar al alumno de Informática a los fundamentos de los Sistemas de Recuperación de la Información; por entender que:

. existe un mercado real en este momento

. es necesario de desarrollar nuevos sistemas que se adapten a la problemática del acceso a la información en la red Internet

. su estudio comparativo con el modelo de bases de datos relacional proporciona una buena oportunidad de fomentar, en los estudiantes de Ciencias de la Computación, el desarrollo de importantes capacidades para su futuro ejercicio profesional.

Se apuntan las líneas metodológicas de la docencia, se ubica el tema dentro del plan de estudios y se evalúa el consumo de recursos necesarios.

Referencias

[1] ACM/IEEE-CS *Joint Curriculum Task force. Cumputing Curricula 1991*

[2] Reales decretos 1460/1990 y 1461/1990 BOE nº 278 de 20 nov 1990 pag 34403-4

[3] C.J. Date; *Introducción a los sistemas de bases de datos*. Addison-Wesley Iberoamericana, 3ª ed. 1986

[4] J.D. Ullman. Data Base Systems. Computer Science Press. 1982

[5] C.J. Date. *Introducción a los sistemas de bases de datos, Vol I*. Addison-Wesley Iberoamericana, 5ª ed. 1993

[6] G. W. Hansen y J. V. Hansen, Diseño y administración de Bases de Datos. Prentice may 1997.

[7] H. F. Korth, A. Silberschatz; *Fundamentos de bases de datos*. McGraw-Hill/Interamericana de España. SA. 3ª ed. 1998

[8] N.Ryan y D. Smith. *Database system Enginiering*. Thomson Computer Press. 1995 ISBN-1-85032-1159

[9] G. Salton: *Introduction to modern Information Retrieval*. McGrawHill (1983) ISSBN 0070544840,pag 3

[10] R. Elmasri y S. B. Navathe; *Sistemas de bases de datos*. Conceptos fundamentales. Addison-Wesley Iberoamericana, SA. 1997. ISBN: 0-201-65370-2.

[11] G.Salton;Automatic text procesing (1989) Addison-Wesley ISBN 0201122278

[12] R.R.Korfhage; *Information Storage and Retrieval*. Willey Computer Publishing (1997)

[13] R.Baeza Yates, B.Ribero-Neto: *Moderm Information Retrieva*l. Addison Wesley (1999)