LA EVOLUCION DE LAS TECNOLOGIAS DE LAS BASES DE DATOS COMO SOPORTE DE LA «INFORMATION RETRIEVAL»: DESDE LOS «THESAURUS» A LAS REDES SEMANTICAS

Por VINCENZO MONACI
Sipe Optimation. Roma

RESUMEN

La evolución de las metodologías de concepción de bases de datos y la innovación tecnológica de sus sistemas de gestión llevan a prever una integración de estos sistemas con los tradicionales sistemas de *information retrieval*. En esta comunicación se examinan los distintos aspectos de esta integración incluso a la luz de las propuestas y de los estudios de los sistemas de la quinta generación.

1. Bases de datos: un escenario para los años noventa

Al principio de los años setenta, cuando los sistemas de bases de datos se afirmaron, primero en el plano científico y experimental, más tarde en el plano productivo y comercial, como la mejor ayuda para la realización de archivos de grandes dimensiones dotados de los requisitos fundamentales de independencia lógico-física, de integración, de protección, de flexibilidad de consulta, etc., apareció una tendencia a la separación entre este tipo de sistemas y sus aplicaciones, por un lado, y los tradicionales sistemas de memorización y recopilación de la información, predominantemente conocidos como sistemas de *information retrieval*, por otro.

En los años sucesivos esta tendencia se fue consolidando a pesar de que, desde un punto de vista puramente técnico, de la expresión

genérica «bancos de datos», con la que se indicaba indiscriminadamente un tipo cualquiera de archivo integrado multiuso, se ha pasado gradualmente a utilizar el término «bases de datos» ampliando su significado más específico. Sucede así que, actualmente, se habla generalmente de bases de datos independientemente del hecho de que se haga referencia, por ejemplo, a bases de datos fácticos, bases de datos de propiedad, bases de datos estadísticos o bases de datos textuales; sin embargo, a todos estos sistemas corresponden profundas diferencias desde el punto de vista de las estructuras y de los lenguajes de gestión, de los métodos de proyecto, de las técnicas de interrogación.

Como nosotros indicamos, los próximos años verán una inversión de esta tendencia. La evolución de las metodologías de concepción de las bases de datos, en efecto, y la innovación tecnológica en sus sistemas de gestión lleva a vislumbrar una integración entre los distintos tipos de bases de datos que está fundamentalmente relacionada con las siguientes características:

— Fusión en el mismo sistema de datos tipológicamente distintos (datos alfanuméricos, informaciones sin enmaquetar, voces, imágenes)
— introducción en los métodos de recopilación de la información de capacidades lógico-deductivas
— desarrollo de metologías de proyección que permitan expresar interdependencias entre los datos no sólo de tipo jerárquico, o a lo sumo racional, sino mucho más complicadas semánticamente
— creación de facetas «amigables» que ayuden al usuario no informático, sino especializado en una materia en particular, a analizar las informaciones memorizadas en las bases de datos o derivables de las mismas.

A largo plazo, estos avances se encuadran en un panorama, el de los sistemas de la *quinta generación,* que en base a los planes de investigación y desarrollo perfilados en el Institute for New Generation Computer Technology, se puede considerar caracterizado por la siguiente constatación. Las calculadoras actuales han sido mejoradas con el fin de realizar rápidamente cálculos de tipo numérico; la evolución de sus aplicaciones ha modificado, sin embargo, las exigencias primeras. Las principales aplicaciones de los ordenadores se referirán cada vez más a la elaboración no numérica de informaciones de naturaleza

multiforme (voces, imágenes, etc.), la gestión de bases de datos, la resolución de problemas.

A medida que evolucionan las aplicaciones de los ordenadores desde la faceta de la actividad del trabajo ejecutivo (contabilidad, gestión de archivos, control de instalaciones) hacia aspectos más elevados del trabajo intelectual, se hace cada vez más necesario el aumento de su capacidad en la dirección de la deducción lógica, de la formulación de planes, de la interacción en lenguaje natural, de la modelización semántica de la realidad: en una palabra, en la dirección de la inteligencia artificial.

La propuesta japonesa para la quinta generación de los sistemas de elaboración pretende responder a estas exigencias apuntando hacia las siguientes direcciones:

I. Al nivel de *hardware,* aprovechando las posibilidades que ofrece la tecnología electrónica VLSI, el objetivo es dotar a la arquitectura de capacidad para ejecutar operaciones deductivas; es decir, que la propuesta consiste en realizar *inference machines* que poseen como actividad elemental de la elaboración la aplicación de una regla de deducción lógica, así como las *computing machines* tienen como actividades elementales la ejecución de instrucciones aritméticas.

II. Al nivel de los lenguajes de programación se pretende desarrollar lenguajes encaminados hacia la *programación lógica* (como, por ejemplo, el lenguaje PROLOG) que permiten describir datos y procedimientos mediante los conceptos lógicos de acciones y reglas de deducción.

III. Al nivel de los sistemas de gestión de las bases de datos el objetivo es el de sustituir tales sistemas por *Knowledge base systems* que permiten representar y elaborar informaciones cada vez más ricas desde el punto de vista semántico y constituyen un apoyo para la resolución de problemas.

IV. Al nivel de las relaciones hombre-máquina, finalmente, se propone la creación de sistemas *multi-media,* con los cuales el usuario pueda utilizar al mismo tiempo voces, imágenes, dibujos.

Aunque la consecución de estos objetivos, destinados a modificar radicalmente las relaciones entre usuarios y sistemas de gestión de las informaciones, resulta por ahora muy lejana, la validez de la dirección señalada es difícilmente discutible.

Por un lado, en efecto, se puede constatar la convergencia esencial hacia esta dirección de los otros proyectos de investigación americanos

y europeos. Particularmente la propuesta europea del proyecto ESPRIT (European Strategic Programme on Research and Development in Information Technology) está orientada también, entre otros objetivos, al estudio de sistemas de gestión de las informaciones basadas en el «conocimiento» y el estudio de sistemas interactivos de interrogación «inteligentes», para usuarios no informáticos.

Por otro lado, independientemente de los avances de la tecnología esperados para el futuro, podemos observar que las tendencias de la investigación en el campo de la metodología de proyectos de las bases de datos se mueven desde hace tiempo en la dirección del enriquecimiento de los aspectos semánticos de las bases de datos y de su intercambio con el usuario. Los avances concretos que se están realizando en este terreno están destinados a influir desde un futuro inmediato sobre los métodos y aplicaciones de dichos sistemas.

2. La información semántica en el proyecto y en la gestión de las bases de datos

El enriquecimiento de los conceptos utilizados en los sistemas de gestión de las bases de datos con los conceptos derivados de los lenguajes de programación y de la inteligencia artificial se remonta al final de los años setenta. En el ámbito de los estudios sobre las metodologías y los lenguajes de concepción de las bases de datos estas interrelaciones se han puesto de manifiesto con claridad. En particular ha surgido de forma crucial la relación entre tipos de datos (tal como se presentan en los lenguajes de programación), dependencias de datos (tal como se definen en las bases de datos) y asociaciones semánticas (tal como se denominan en el ámbito de la inteligencia artificial). En efecto, mientras los actuales sistemas de gestión de las bases de datos prevén la representación de asociaciones entre los datos, particularmente elementales, es decir, por ejemplo, de tipo jerárquico o relacional, tipos de asociaciones más complejas están empezando a afirmarse (por ejemplo, en el ámbito del modelo «entidad-asociación» o de su ampliación realizada en el objetivo DATAID del Proyecto Finalizado Informática del CNR) y están empezando a presentarse en el mercado lenguajes para bases de datos que disponen de una riqueza semántica todavía más amplia, derivada de los conceptos de inteligencia artificial (como el lenguaje TAXIS o el lenguaje GALILEO, desarrollado también en el proyecto DATAID).

El instrumento conceptual en que se basan estos avances son las denominadas *redes semánticas,* estructuras basadas esencialmente en objetos y asociaciones binarias de varios tipos.

Las redes semánticas, nacidas en el ámbito de los primeros estudios sobre la inteligencia artificial y sobre el lenguaje natural fueron introducidas a mitad de los años sesenta y han sido recientemente retomadas en el ámbito de los sistemas de bases de datos.

Como instrumentos de representación conceptual de la realidad, pueden relacionarse con el modelo CODASYL aunque representan un nivel de abstracción muy superior. Entre los tipos de asociaciones más importantes podemos recordar la clasificación, la agregación y la generalización, esta última correspondiente a la relación de tipo jerárquico que rige, por ejemplo, en un «thesaurus». La riqueza conceptual de las redes semánticas hace prever un papel cada vez más importante de las mismas en la evolución de las «bases de datos» hacia las «bases de conocimiento».

También las habilidades deductivas que están presentes desde hace tiempo en las investigaciones sobre inteligencia artificial orientadas hacia el desarrollo de sistemas de gestión de bases de conocimiento están haciendo su aparición en el campo de las bases de datos.

Una *base de datos deductiva* es una base de datos en la cual nuevas informaciones no memorizadas directamente pueden deducirse de otras informaciones explícitamente introducidas.

Este tipo de avance de las bases de datos resulta particularmente apto para la realización de *sistemas expertos.* En estos sistemas, en efecto, más que la disponibilidad de grandes cantidades de datos (los sistemas expertos deberían funcionar también sobre ordenadores personales) resultan necesarias:

— Capacidad de deducir informaciones en base a los datos disponibles y a las leyes generales de una determinada materia
— capacidad de formular escenarios para la verificación de la validez de hipótesis y de modelos
— capacidad de facilitar el diálogo, tanto en fase de proyecto como en fase de consulta, con un usuario no informático, aunque especializado en la materia en cuestión.

Esto es lo que sucede, por ejemplo, en los sistemas expertos realizados para facilitar el diagnóstico médico, para la formulación de hipótesis en química orgánica, para facilitar las decisiones en el ámbito legal, notarial y jurídico.

A la luz de los avances realizados en el enriquecimiento lógico-semántico de las bases de datos resulta siempre realista, por ejemplo, prever (y dirigirse hacia) un desarrollo del diccionario de los datos que, de simple instrumento de catalogación y de clasificación de las informaciones disponibles en una base de datos, se convierte en el centro de referencia de toda la interacción usuario-sistema, ya sea en el momento de la concepción (definición de datos, procedimiento, sucesos y de su interrelación que constituyen la red semántica) como en el momento de la consulta y de la investigación o derivación de informaciones (realizable mediante una exploración de los caracteres contenidos en la red y sus asociaciones).

3. Adaptación y reconfiguración en las bases de datos

Otro aspecto que se deriva de la evolución lógico-semántica de las bases de datos se refiere a la *interfaz* de usuario y la posibilidad de autoadaptación de las bases de datos.

Tradicionalmente, la fase de proyecto de una base de datos en sus diversos aspectos, conceptuales, lógicos y físicos, es una fase que se desarrolla esencialmente en modalidad no interactiva. Partiendo de entrevistas y de módulos, las exigencias informativas del usuario quedan poco a poco formuladas y codificadas siguiendo los pasos de un método manual, eventualmente apoyado por instrumentos, o bien por medio del uso de un auténtico y verdadero lenguaje de programación.

Las consideraciones desarrolladas anteriormente nos llevan a formular la hipótesis de un panorama muy distinto para el futuro. La mayor riqueza de conceptos que una red semántica puede alcanzar y la mayor flexibilidad en la determinación de las asociaciones entre los datos, nos inducen a ver el proyecto de una base de datos (y en el futuro una base de conocimiento) como un proceso gradual durante el cual el usuario no informático introduce poco a poco las especificaciones lógico-formales de la realidad que pretende representar, en una relación interactiva con el sistema que lo guía en la construcción del modelo. Esto significa que las entidades, las asociaciones, las propiedades que resultan progresivamente definidas por el usuario deben verificarse según su consistencia y eventualmente modificarse sucesivamente. Además, conceptos no definidos durante la fase inicial del proyecto pueden surgir durante las fases sucesivas de gestión y consulta. Es, por tanto, necesario que el diccionario de datos sea visto en general como una estructura que evoluciona y «capta» elementos nuevos que poco a

poco van integrando la visión conceptual de la realidad proporcionada inicialmente por el usuario.

Un proceso análogo de «aprendizaje» debe desarrollarse para permitir la adaptación de las prestaciones de la base de datos a las exigencias del usuario. También desde este punto de vista la evolución de los sistemas de gestión de las bases de datos llevará a la superación de la tradicional dicotomía entre sistemas eficaces de consulta, pero esencialmente estáticos y de difícil actualización y sistemas dinámicos y flexibles en la actualización y lentos en la consulta. Un sistema de gestión de bases de datos «inteligente» podrá, en efecto, autoorganizarse sobre la base de las frecuencias de consulta a través de las distintas vías de acceso.

4. Conclusiones

Las innovaciones de los sistemas de gestión de bases de datos que desde hace tiempo se persiguen en el ámbito de algunos proyectos de investigación italianos e internacionales, y que presumiblemente extraerán ulteriores incentivos del avance de los proyectos sobre los calculadores de la quinta generación, están esencialmente orientadas a dotar a los propios sistemas de capacidad de representación y gestión de la información semántica. Los sistemas de bases de datos «inteligentes» podrán:

— Integrar los datos heterogéneos como textos, voces, imágenes
— utilizar la descripción de la realidad y de sus propiedades para asistir al usuario en la deducción de sucesos no disponibles directamente
— autoadaptarse a la evolución conceptual o física de las exigencias del usuario.

Desde esta perspectiva, las distinciones tradicionales entre sistemas de *information retrieval* y sistemas de *data base management* está destinada a ser superada, ya que tanto los unos como los otros se englobarán en los más complejos sistemas de gestión del conocimiento.