BANCOS DE DATOS GEOGRAFICOS Y MODELO RELACIONAL

Hector Daniel Castro
Porto Alegre - *Brasil*

INTRODUCCION

Entre las aplicaciones no convencionales de los bancos de datos se encuentran aquellas dedicadas a los esfuerzos de planificación, en las que el acúmulo de datos organizados geográficamente juega un papel importante. Si bien estos datos pueden ser de tipos muy diversos: climáticos, hídricos, económicos, demógraficos, etc., su característica común es que de alguna manera están relacionados con una ubicación geográfica. Las organizaciones propuestas para los bancos de datos geográficos utilizan en su mayoría el modelo de datos relacional. Se analizarán las maneras de representar los datos geográficos y tres implementaciones de bancos de datos geográficos. Esta presentación forma parte de un proyecto orientado a obtener un esquema conceptual de un banco de datos geográficos.

REPRESENTACIONES DE DATOS GEOGRAFICOS [1]

Conforme fue explicado anteriormente, en un banco de datos geográficos existen, en principio, dos tipos de datos: aquellos que representan propiedades específicas del tema siendo estudiado (por ejemplo, número de personas o valores económicos), que serán denominados "datos simbólicos"; y aquellos que representan la ubicación espacial (geográfica) a que están referenciados los primeros. Estos ultimos serán llamados "datos geográficos". Es claro que entre los datos simbólicos pueden existir sub-clasificaciones, o incluso pueden no aparecer por completo, dependiendo de la aplicación. En cambio, los datos geográficos deben estar siempre presentes, y para ellos debe usarse en todo el banco de datos una representación uniforme.

Para objetivar, vamos a dar un ejemplo. Supongamos un banco de datos destinado a describir la red vial de una región. Si consideramos que dicha red esta compuesta por rutas que unen ciudades y nuestro interés fuera hallar la manera de ir de una ciudad a otra, bastaría con representar la ciudad de origen y la ciudad de destino de cada ruta, y tal vez las intersecciones de las rutas. Aquí sólo existe información geográfica, si bien que de manera implícita a través de la ubicación espacial de cada ciudad, que suponemos conocida, y que, obviamente, no varía con el tiempo. Si, además, estuviéramos interesados en hallar la mejor ruta, basándonos en criterios como tipo de la calzada o volumen de tránsito, deberíamos incluir esos datos en

nuestro banco. Estos últimos datos pertenecen a los que llamamos simbólicos. Los datos simbólicos pueden ser de dos tipos: numéricos (volumen de tránsito) o identificatorios (tipo de calzada). Vemos que todos los datos están relacionados espacialmente, aunque sea en forma indirecta: el tipo de calzada se refiere a una ruta, y ésta es ubicada por sus puntos extremos.

Cabe aclarar que, en nuestro concepto, un banco de datos como el descripto no sería propiamente un banco de datos geográficos. Para ser tal, falta un concepto importante, que es el de medida. En el banco de datos presentado, no podemos averiguar la cantidad de kilómetros que una ruta emplea para unir dos ciudades. Para conseguir eso deberíamos agregar explícitamente la ubicación de cada ciudad, mediante sus coordenadas geográficas, si suponemos que las rutas son rectas. Si no son rectas, deberíamos descomponerlas en segmentos rectos. Nótese que no bastaría con agregar esa información a los datos simbólicos, porque queremos asimilar el uso del banco de datos geográficos al uso de un mapa. Es decir, deberían poder realizarse medidas encima del banco de datos de la misma manera como son hechas encima del mapa.

Existen dos maneras usadas comunmente para representar datos geográficos, que permiten realizar mediciones. El primer método, que llamaremos "de reja", consiste en cuadricular el área cubierta por los datos, y ubicar espacialmente cada elemento de área por el número de fila y el número de columna, y las coordenadas de una de las esquinas de la cuadrícula. Los datos simbólicos son entonces referidos a estos elementos de área. En el caso de variables del tipo $z = f(x,y)$, asociaremos con cada elemento de área un número o un identificador. Este es el tipo de datos para el que es más útil el método de reja. Si queremos representar datos lineales, como por ejemplo una vía de ferrocarril, deberíamos recurrir al expediente de indicar su presencia en cada uno de los elementos de área que ella atraviesa, y su ausencia (por "default") en todos los otros. Es de hacer notar que usando este método es fácil calcular la superfície relacionada con un atributo, simplemente multiplicando la superfície de un elemento de área por la cantidad de elementos con ese atributo. Igualmente fácil es calcular la distancia entre dos elementos de área. Una variante de este método es el llamado método "raster", donde cada elemento de área es asimilado a un punto de una imagen, que puede representar sólo un atributo. No se admiten aquí valores por "default" y todo punto tiene un valor. Otra variante es el método llamado "quadtree", donde la superfície de cada elemento de área es variable conforme la necesidad de representar con mayor o menor detalle, una porción del área total.

El segundo método importante, que llamaremos "vectorial", es más adecuado para representar objetos de características lineales. Los mismos son descritos usando listas de coordenadas correspondientes a los puntos de início y fin de los segmentos que componen el objeto. Un punto es representado por sus coordenadas, y un área a través de sus fronteras. Este método fue usado originalmente para representar datos urbanos, y su aplicación principal es la representación de mapas compuestos por polígonos disjuntos. Para este tipo de tarea hace un uso más eficiente de la memoria que el método anterior. Es también posible hacer, a traves de algoritmos, cálculos de la superfície, centro de gravedad y perímetro de un polígono, y determinar si un punto se encuentra dentro de un polígono.

EL MODELO RELACIONAL [2]

En el modelo relacional los datos son organizados usando relaciones. Se puede definir una relación de la siguiente manera: Sea una colección de conjuntos D_1 , D_2, ..., D_n(no necesariamente distintos), R es una relación de los conjuntos D_1, D_2, ..., D_n si es un subconjunto del producto cartesiano D_1 x D_2 x .... x D_n . El valor n es llamado el grado de la relación. Los conjuntos D_i son llamados domínios y los elementos de R, tuplas. El número de tuplas en una relación es la cardinalidad de la misma. Un atributo es el conjunto de valores extraídos de un domínio, que es usado en una relación. Se dice que una relación está normalizada o en primera forma normal cuando cada valor de un atributo es elemental, es decir, no admite descomposición. Un atributo de una relación es llamado clave primaria de esa relación, cuando los valores de ese atributo son diferentes para cada tupla de la misma. Representaremos una relación de la siguiente manera:

nombrerelación (Atributo1, Atributo2, ..., AtributoN)

Si una relación no tiene una clave primaria formada por un único atributo, siempre puede ser obtenida una clave primaria concatenando dos o más atributos. Si hay más de un atributo o combinación de atributos con la propiedad de poder ser usados como identificadores de tuplas, reciben en conjunto el nombre de claves candidatas. Aquellas claves candidatas que no son la clave primaria, son llamadas claves alternativas. Si un atributo de una relación es la clave primaria de otra, es llamado clave extranjera. Forman parte del modelo relacional las restricciones de integridad. Estas dicen que, en una relación, los valores de la clave primaria no pueden ser nulos (integridad de entidad) y que en el caso

de una relación tener una clave extranjera, cada valor de ese atributo tiene que aparecer como clave primaria de una tupla en una segunda relación (integridad referencial). Un banco de datos relacional es, entonces, un conjunto de relaciones normalizadas que cumplen las condiciones especificadas encima, con respecto a las claves y a la integridad.

IMAID [3]

Es un sistema que integra análisis de imágenes y gerencia de banco de datos. Fue desarrollado en la Universidad de Purdue con el propósito de almacenar imágenes de los satélites LANDSAT. Un mapa extraído de fotos satelitárias y compuesto por trazos rectilíneos (dibujo lineal) puede ser convertido en una relación con cuatro atributos especificando las coordenadas (x,y) de los puntos extremos de cada segmento. Pueden agregarse atributos adicionales si hubiera necesidad. Un punto es representado como un segmento de línea con sus dos extremos coincidentes. Una línea es representada por un conjunto de segmentos. Una región es representada por su frontera. Por ejemplo, en un mapa vial extraído de fotos satelitárias podríamos tener datos de rutas y ciudades. Cada foto, ruta y ciudad recibiría un número identificatorio: Nfoto, Nruta y Nciudad, respectivamente. Para cada segmento indicaríamos sus puntos extremos dentro de la foto y una ciudad sería tratada como una región. Ademas deseamos saber qué rutas y qué ciudades hay en cada foto, y la ubicación y otros datos de cada foto. Tendríamos entonces las siguientes relaciones:

rutas (Nfoto, Nruta, X1, Y1, X2, Y2)
nombreruta (Nfoto, Nruta, Nombre)
posición (Nfoto, Xtam, Ytam, Xcen, Ycen, Archivo)
ciudades (Nfoto, Nciudad, X1, Y1, X2, Y2)
nombreciudad (Nfoto, Nciudad, Nombre)

En la relación 'rutas', como puede haber más de un segmento para cada ruta, conseguimos la condición de unicidad de clave primaria, concatenando todos los atributos. Lo mismo ocurre con la relación 'ciudades'. Las relaciones 'nombreruta' y 'nombreciudad' tienen como clave primaria Nfoto-Nruta y Nfoto-Nciudad, dado que una ruta y una ciudad no pueden aparecer más de una vez en una foto. La relación 'posición' tiene como clave primaria Nfoto. Esto hace que el atributo Nfoto sea clave extranjera en las relaciones 'rutas', 'nombreruta', 'ciudades' y 'nombreciudad'. En las relaciones 'nombreruta' y 'nombreciudad' podemos tener la clave alternativa Nfoto-Nombre, si suponemos unicidad de los nombres.

GEO-QUEL [4]

Es una extensión al sistema de gerencia de banco de datos INGRES, desarrollado en la Universidad de Berkeley. Un mapa consta de una sola relación, donde cada tupla corresponde a un segmento de recta cuyos extremos están dados por los atributos X1, Y1, X2, Y2. Un punto es considerado como un segmento con extremos coincidentes. Regiones y líneas son considerados como agregaciones de segmentos y reciben un identificador (Idlínea e Idregión). Existe un atributo (Tipo) que indica si una tupla corresponde a un punto, un segmento aislado, una línea, una región o un punto que es el centro de una región. Además existen atributos destinados a especificar la manera de visualización de los datos. Esa unica relación respondería al esquema

mapa (X1, Y1, X2, Y2, Tipo, Vis1, Vis2, Idlínea, Idregión)

Vis1 y Vis2 son atributos de visualización. Otros atributos podrían agregarse.

Dado que un mapa es considerado como una colección de polígonos disjuntos, un segmento puede pertenecer a dos polígonos, o sea que las coordenadas no pueden ser consideradas como clave primaria. Una tupla siempre tiene un identificador de línea, pero puede tener un identificador nulo de región, asi que la combinación X1-Y1-X2-Y2-Idlínea sería la clave primaria de esta relación. No existen claves extranjeras al existir una sola relación.

DIMAP [5]

Este sistema combina un sistema de gerencia de banco de datos con un sistema de almacenamiento de imágenes y fue desarrollado en la Universidad de Illinois. Un banco de datos es un conjunto de mapas organizado jerárquicamente. Un mapa es un conjunto de relaciones, cada una correspondiendo a un tipo de objetos pictóricos. De cada una de estas relaciones se obtienen, por restricción, las relaciones correspondientes a un cuadro, que es la mínima unidad visualizable. Cada objeto pictórico corresponde a una tupla. Por ejemplo el mapa de una región podría estar formado por tres relaciones: 'ciudades', 'rutas' y 'ríos'. En un nivel inferior de la jerarquía, una ciudad incluída en esa región podría tener un mapa compuesto de dos relaciones: 'distritos' y 'población'. Para cada mapa existe una relación especial que explicita las relaciones componentes y el orden jerárquico de las mismas. Tambien existe para cada mapa una relación especial que indica el nombre de un

programa que grafica un cuadro de esa relación. En DIMAP los puntos son representados por sus coordenadas, una línea por sus extremos y un área es representada usando el metodo "raster", es decir, por sus puntos constituyentes.

Un mapa de zonas de vegetación sería representado por dos relaciones como:

vegetación (X, Y, Clase)
clasevegetación (Clase, Descripción)

En la relación 'vegetación' la clave primaria está formada por las coordenadas, mientras que Clase es una clave extranjera. Un mapa de puentes sería definido por una relación del tipo punto como:

puente (Nombre, X, Y, Tipopuente, ...)

Un mapa de rutas sería definido por una relación del tipo línea como

ruta (Nombre, X1, Y1, X2, Y2, Claseruta, ...)

En estas relaciones el nombre y/o las coordenadas forman la clave primaria. Claseruta y Tipopuente son claves extranjeras.

CONCLUSIONES

La primera conclusión que puede extraerse de la consideración de los sistemas expuestos es que, en el proyecto de un banco de datos geográficos, debe dedicarse gran atención a los aspectos de modelización, dado que la diversidad de enfoques encontrada indica que la complejidad de la aplicación es alta.

Si bien el método vectorial es más ampliamente usado, existen variantes en lo que respecta a la manera de agrupar en relaciones las coordenadas que representan los objetos y en cuanto al grado de explicitamiento con que éstos aparecen en el banco de datos. Por ejemplo, un punto puede constar de un par de coordenadas (x,y) o de dos pares idénticos de coordenadas, y una región puede aparecer como una entidad por derecho propio, es decir, en una relación de regiones, o sólo implícitamente, a través de un atributo con un valor común en todas las tuplas de segmentos que forman la frontera de la región.

El criterio de minimizar el número de relaciones influye en la fácilidad con que pueden ser agregados atributos a las entidades y hace que el esquema del banco de

datos sea bastante rígido. Un caso claro de esto es el sistema GEO-QUEL, donde un atributo da el significado de cada tupla en una relación única, sin permitir una caracterización diferenciada de cada objeto.

Otro problema de modelado está indicado por la aparición de claves primarias compuestas por varios atributos, en algunos casos abarcando toda la tupla, sin que esto se origine en relacionamientos entre entidades, como es lo normal.

La segunda conclusión que surje del análisis es que el modelo relacional, como fue expuesto, no permite expresar a través de sus mecanismos intrínsecos, dos características importantes de los datos geográficos. Estas son la agregación de objetos (v.g. segmentos para formar líneas) y el ordenamiento jerárquico dado por la inclusión de un objeto en otro (v.g. una ciudad en una región).

Las mismas son representadas mediante el uso de claves extranjeras, que en princípio deberían resolver problemas de dependencias funcionales, o sino recurriendo a relaciones especiales como la relación POT del sistema DIMAP, con atributos que tienen como valores nombres de relaciones, lo que no está permitido en el modelo relacional, ya que una relación es un valor compuesto.

En un proyecto en curso se están abordando estos temas con el objetivo de lograr una representación de los datos geográficos independiente de cualquier modelo de banco de datos y de determinar la utilidad de un modelo relacional expandido para la obtención de un esquema de bancos de datos geográficos.

REFERENCIAS

[1] Castro, H. D. "Concepts of geographic information systems". UFRGS-CPGCC, Porto Alegre, 1986.

[2] Date, C. J. "An introduction to database systems", Addison-Wesley, Reading, 1983.

[3] Chang, N. S. y Fu, K. S. "A relational database system for images" en "Pictorial information systems", Springer-Verlag, Berlin, 1980.

[4] Berman, R.R. y Stonebreaker, M. "GEO-QUEL: A system for the manipulation and display of geographic data". Computer Graphics 11(2), Summer 1977.

[5] Chang, S. K., Lin, B. S. y Walser, R. "A generalized zooming technique for pictorial database systems" en "Pictorial information systems", Springer-Verlag, Berlin, 1980.