Indexación y Búsquedas en Bases de Datos

Anabella De Battista, Andrés Pascal
Juan Pablo Nuñez, Soledad Retamar
Departamento Ingeniería en Sistemas de Información
Fac. Reg. Concepción del Uruguay
Universidad Tecnológica Nacional
Entre Ríos, Argentina
{pascalj, debattistaa, nunezjp, retamars}@frcu.utn.edu.ar

Norma Edith Herrera
Departamento de Informática
Univ. Nac. de San Luis
San Luis, Argentina
nherrera@unsl.edu.ar

Gilberto Gutierrez
Facultad de Ciencias Empresariales
Universidad del Bio-Bio
Chillán, Chile
ggutierr@ubiobio.cl

Resumen

En la actualidad se utilizan nuevos modelos de bases de datos en los que no es aplicable el concepto de búsqueda exacta que se emplea en las bases de datos tradicionales. La causa principal es que estos nuevos modelos de bases de datos permiten gestionar tipos de datos con características que difieren de los datos clásicos: no es posible establecer un orden entre ellos y, en general, no pueden estructurarse por lo que no pueden almacenarse en registros ni campos. Estos tipo de datos pueden ser sonidos, imágenes, videos, entre otros, y para realizar consultas sobre los mismos se emplea el concepto de búsqueda por similitud. También resulta de interés en algunas aplicaciones poder consultar estados pasados de la base de datos y no solamente el estado actual. A raíz de estas situaciones, se plantean nuevas estrategias para el almacenamiento y la consulta en las bases de datos, entre los más conocidos podemos mencionar los modelos de espacios métricos, métrico-temporal, espacial, temporal, espacio-temporal. En este artículo se presentan los tópicos de interés del *Grupo de Investigación en Bases de Datos (GIBD)*: modelado de objetos no estructurados y procesamiento eficiente de consultas sobre estos tipos de datos.

Palabras Claves: Bases de Datos Espaciales, Bases de Datos Temporales, Bases de Datos Espacio-Temporales, Espacios Métricos, Índices, Espacios Métrico-Temporales.

1. Contexto

El presente trabajo se desarrolla en el ámbito del proyecto *Procesamiento eficiente de consultas en nuevos Modelos de Bases de Datos* (PID 25-D059) del Grupo de Investigación en Bases de Datos, perteneciente al Departamento Ingeniería en Sistemas de Información de la

Universidad Tecnológica Nacional, F. R. Concepción del Uruguay.

2. Introducción

En la actualidad resulta necesario gestionar nuevos modelos de bases de datos que permitan almacenar y consultar de manera eficiente ciertos tipos de datos no estructurados como imágenes (médicas, satelitales, huellas dactilares, etc), audio, textos, videos, entre otros. En este ámbito resulta necesario implementar nuevas estrategias de almacenamiento y de consulta, teniendo en cuenta que generalmente no es posible estructurar este tipo de datos, por lo que no puede almacenarse la información en registros y campos. Por tales motivos, es que no puede aplicarse el concepto de búsqueda exacta utilizado en las bases de datos tradicionales, por lo que índices tales como el *B*-Tree* no son utilizables para realizar búsquedas de manera eficiente. En estos casos se aplica el concepto de búsquedas por similitud, en las que dos objetos se comparan mediante una función de distancia que indica el grado de similitud que existe entre ellos. El conjunto de objetos se denomina espacio métrico. Adicionalmente, existen aplicaciones en las que se requiere almacenar mas de un estado de la base de datos, y no solamente el estado actual. Por lo que esta situación debe tenerse en cuenta tanto en las estrategias de almacenamiento como en las de búsquedas a implementar.

Como respuesta a estos nuevos requerimientos se han generado los nuevos modelos que se describen a continuación.

Para el almacenamiento de objetos no estructurados se ha propuesto el modelo de *Espacios Métricos* , que permite realizar consultas por similitud eficientemente. Este tipo de consultas utiliza funciones de distancia para determinar el grado de similitud entre los objetos de la base de datos y el objeto que se consulta. Un *Espacio Métrico* se define como un par (U, d) donde U es el universo de objetos válidos del espacio y $d : U \times U \longrightarrow R^+$ es una función métrica que se define entre los elementos de U y que permite medir su similitud (a menor distancia más cercanos o similares son los objetos). Llamaremos base de datos a cualquier subconjunto finito $X \subseteq U$ cuya cardinalidad es $|X| = n$. La función d cumple con las propiedades características de una función métrica: $\forall x, y \in U,\ d(x, y) \geq 0$ (positividad); $\forall x, y \in U,\ d(x, y) = d(y, x)$ (simetría); $\forall x \in U,\ d(x, x) = 0$ (reflexividad) y $\forall x, y, z \in U,\ d(x, y) \leq d(x, z) + d(z, y)$ (desigualdad triangular). En base a este modelo se han desarrollado índices especiales que aumentan la velocidad de respuesta de las búsquedas por similitud.

Para poder gestionar objetos con alguna referencia espacial se han propuesto las *Bases de Datos Espaciales* [4] permiten . Un dato espacial puede ser en su forma más simple un punto, una polilínea o un polígono. La persistencia de estos tipos de datos espaciales se basa no sólo en el valor de ciertos atributos, sino también en la ubicación espacial del objeto. Por ejemplo, podría resultar de interés obtener los terrenos geográficamente adyacentes a uno dado, o encontrar todos los hospitales cercanos a una determinada ruta. Existen muchas aplicaciones para el modelo de bases de datos espaciales; una de las más destacadas son los sistemas de información geográfica (SIG), que realizan el procesamiento de datos geográficos y que almacenan la geometría y los atributos de datos con algún tipo de georreferencia, es decir, situados en la superficie de la tierra y representados bajo una proyección cartográfica. Uno de sus objetivos es resolver problemas complejos de planificación y gestión.

A fin de poder asociar tiempos de vigencia a objetos almacenados han surgido las *Bases de Datos Temporales*, que permiten manejar internamente una o más dimensiones temporales. Existen tres clases de bases de datos temporales según la forma en que manejan el tiempo: (a) de tiempo transaccional (transaction

time), donde el tiempo se registra de acuerdo al orden en que se procesan las transacciones; (b) de tiempo válido (valid time), que almacenan el momento en que el hecho ocurrió en la realidad, que puede no coincidir con el momento de su registro; y (c) bitemporales, que integran la dimensión transaccional y la dimensión vigente a través del versionado de los estados. En las consultas se requiere conocer el comportamiento de algún objeto en algún instante dado o durante un intervalo de tiempo determinado. Por ejemplo una consulta temporal podría ser *recuperar la evolución del sueldo de un empleado en un intervalo de tiempo dado*, o *encontrar todos los empleados que tenían cierta categoría en una fecha dada.*

Para abordar aquellas aplicaciones que requieren la resolución de consultas que involucran más de un aspecto de los antes mencionados se plantean combinaciones de estos modelos de bases de datos . Así han surgido los modelos *Espacio-Temporal* y *Métrico-Temporal.*

Las *Bases de Datos Espacio-Temporales* permiten gestionar la naturaleza dinámica de los objetos espaciales [9]. Las consultas a resolver en este tipo de bases de datos pueden incluir referencias espaciales, tales como posición, intersección, inclusión o superposición, y temporales, tanto respecto al pasado o presente como predicciones del tiempo futuro. Por ejemplo, nos puede interesar saber cuál es la máxima velocidad alcanzada por un objeto en un intervalo de tiempo, o recuperar los objetos que cruzaron una cierta área en un instante de tiempo dado o incluso los que pasarán por un punto en el futuro, considerando su dirección. Constituyen el ámbito de aplicación de este modelo de bases de datos las aplicaciones de predicción climática, control de tráfico terrestre o aéreo, aspectos sociales (demografía, salud) y multimedia.

El *Modelo Métrico-Temporal* surge ante la necesidad de aplicaciones donde resulta de interés realizar búsquedas por similitud teniendo en cuenta también la componente temporal. En este modelo se puede trabajar con objetos no estructurados con tiempos de vigencia asociados y realizar consultas por similitud y por tiempo en forma simultánea. Formalmente un *Espacio Métrico-Temporal* es un par (U, d), donde $U = O \times N \times N$, y la función d es de la forma $d : O \times O \rightarrow R^+$. Cada elemento $u \in U$ es una triupla (obj, t_i, t_f), donde obj es un objeto (por ejemplo, una imagen, sonido, cadena, etc) y $[t_i, t_f]$ es el intervalo de vigencia de obj. La función de distancia d, que mide la similitud entre dos objetos, cumple con las propiedades de una métrica (positividad, simetría, reflexividad y desigualdad triangular). Una *consulta métrico-temporal* por rango se define como una 4-upla $(q, r, t_{iq}, t_{fq})_d$, tal que $(q, r, t_{iq}, t_{fq})_d = \{o/(o, t_{io}, t_{fo}) \in X \wedge d(q, o) \leq r \wedge (t_{io} \leq t_{fq}) \wedge (t_{iq} \leq t_{fo})\}$.

3. Líneas de Investigación

La línea de trabajo principal de nuestro grupo es el estudio de métodos de acceso, procesamiento de consultas y aplicaciones de bases de datos no tradicionales, principalmente de los modelos métrico-temporal y espacio-temporal. Damos a continuación una descripción de las líneas de investigación que actualmente estamos desarrollando.

3.1. Implementación de Indices Métrico-Temporales en memoria secundaria

Para el tratamiento de objetos métrico-temporales se han propuesto los índices *FHQT-Temporal* [8], *Historical-FHQT* [2], *Event-FHQT* [7] y *Pivot-FHQT* [3], que toman como base el índice para espacios métricos Fixed Height Queries Tree[1], que utiliza funciones de distancia discretas. Se han propuesto variantes que permiten tanto funciones discretas como continuas: *FHQT$^+$-Temporal* y

Event-FHQT$^+$.

Los índices desarrollados hasta el momento han sido evaluados empíricamente con lotes generados a partir de imágenes del sitio *SISAP* (http://www.sisap.org), asociando a cada imagen un intervalo de vigencia.

En dichas pruebas se supone que tanto los datos como el índice pueden mantenerse en memoria principal, pero como estas bases de datos son de gran tamaño actualmente se está trabajando en la implementación de dichos índices en disco.

3.2. Aplicaciones de Bases de Datos Espaciales, Espacio-Temporales y Sistemas de Información Geográfica

En el marco de este proyecto se han firmado convenios de colaboración con otras instituciones y grupos de investigación con el fin de prestar servicios relacionados a la temática del grupo. Actualmente se está colaborando con el Grupo de Estudios de Calidad y Medio Ambiente de la Regional Concepción del Uruguay de la UTN en la implementación de un Sistema de información Geográfica para el Municipio de la ciudad de Urdinarrain (Entre Ríos), a fin de obtener una herramienta de planificación para el sector comercial de dicha localidad. Se desarrolló una aplicación que permitirá gestionar los datos obtenidos a partir del relevamiento de comercios, asociando cada uno de ellos a su posición geográfica, para permitir un posterior análisis de distribución de comercios por rubro, zona geográfica y otros factores de interés para el caso particular. Por otra parte, con la Facultad de Ciencias de la Salud de la Univ. Nac. de Entre Ríos (FCS-UNER) se estableció un convenio para el desarrollo y mantenimiento de un servidor de mapas interactivo en el que se visualizan datos georreferenciados resultantes de diversos proyectos de investigación de dicha institución. En una segunda etapa se está elaborando una capa geográfica a partir de la base de datos de alumnos ingresantes que represente la ciudad de origen de los mismos y las carreras de grado que brinda dicha institución universitaria. Dos integrantes del proyecto colaboran además con un investigador de FCS-UNER que está desarrollando su tesis doctoral en Geografía, en el análisis de datos obtenidos de registros hospitalarios, que permitirán elaborar conclusiones sobre la accesibilidad geográfica de la población de la provincia de Entre Ríos a los centros de salud [6, 5].

4. Resultados Esperados

Se espera contar con métodos eficientes, tanto en memoria principal como en memoria secundaria, para el procesamiento de consultas en el ámbito de bases de datos no tradicionales. Esto incluye el diseño de índices, la definición de funciones de distancias adecuadas a la problemática tratada, la definición de nuevas consultas que sean de interés y el desarrollo de aplicaciones en ámbitos reales de uso de los métodos desarrollados. Además se continuarán realizando actividades de extensión en el marco de convenios con otras instituciones a fin de difundir las tareas realizadas por el grupo de investigación.

5. Formación de Recursos Humanos

El trabajo desarrollado hasta el momento forma parte del desarrollo de dos Tesis de Maestría en Ciencias de la Computación con orientación en Bases de Datos que se dicta en la Fac. Reg. Concepción del Uruguay de la UTN (MCC-FRCU-UTN), que ya han sido defendidas y aprobadas. Dos de los integrantes del proyecto se encuentran realizando dicho posgrado, uno de ellos ha obtenido una beca de la universidad y está comenzan-

do su trabajo de tesis en la temática bases de datos espacio-temporales, trabajando específicamente con agrodatos. La codirectora del proyecto es directora de dicho becario y además codirige una tesis de Maestría en Ingeniería en Sistemas de Información de la Fac. Regional Córdoba de la UTN. La Directora y la codirectora del proyecto han dictado cursos de posgrado en la MCC-FRCU-UTN. Uno de los integrantes del grupo está desarrollando su Tesis Doctoral sobre la temática de indexación en memoria secundaria de bases de datos textuales, tema íntimamente relacionado a las líneas de estudio de este grupo. El grupo cuenta en la actualidad con dos becarios alumnos de la carrera Ingeniería en Sistemas de Información que se están formando en estas temáticas. Se han desarrollado hasta la fecha siete trabajos finales de dicha carrera de grado en el marco del proyecto.

Referencias

[1] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In *Proc. 5th Combinatorial Pattern Matching (CPM94)*, LNCS 807, pages 198–212, 1994.

[2] A. De Battista, A. Pascal, G. Gutierrez, and N. Herrera. Un nuevo indice metrico-temporal: el historical fhqt. In *Actas del XIII Congreso Argentino de Ciencias de la Computacion*, Corrientes, Agentina, 2007.

[3] A. De Battista, A. Pascal, N. Herrera, and G. Gutierrez. Metric-temporal access methods. *Journal of Computer Science & Technology,*, 10(2):54–60, 2010.

[4] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, Santa Clara, CA, USA, 3rd ed. edition, 2008.

[5] Gagliardi E. Taranilla M. Hernández Peñalver G. Loyola R. Casanova C., Dorzán M. Una aplicación de workforce managment utilizando geometría computacional y bases de datos espacio temporales. *CACIC 2010*, 1(1), 2010.

[6] Dorzán M. Gagliardi O. y Taranilla M. Guasch M., Piergallini M. Bases de datos espacio-temporales aplicadas al análisis y seguimiento de focos epidémicos. *CACIC 2014*, 1(1), 2014.

[7] A. Pascal, A. De Battista, G. Gutierrez, and N. Herrera. Indice metrico-temporal event-fhqt. In *Actas del XIIII Congreso Argentino de Ciencias de la Computacion*, La Rioja, Argentina, 2008.

[8] A. Pascal, De Battista, G. Gutierrez, and N. Herrera. Procesamiento de consultas metrico-temporales. In *XXIII Conferencia Latinoamericana de Informatica*, pages 133–144, San Jose de Costa Rica, 2007.

[9] Gilberto Antonio Gutierrez Retamal. *Metodos de Acceso y Procesamiento de Consultas Espacio-Temporales*. PhD thesis, Universidad de Chile, 2007.