Consultas en nuevos modelos de Bases de Datos

Andrés Pascal, Anabella De Battista ,
Juan Pablo Nunez, Soledad Retamar, Melisa Argüello, Christian Saliwonczyk
Departamento Ingeniería en Sistemas de Información
Fac. Reg. Concepción del Uruguay
Universidad Tecnológica Nacional
Entre Ríos, Argentina
{pascalj, debattistaa, nunezjp, retamars, arguellom, saliwonczykc}@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

Por la capacidad de almacenar datos estructurados que poseen las bases de datos tradicionales se aplica en este modelo el concepto de búsqueda exacta, es decir consultas por exactitud o por rango de valores suceptibles de ser ordenados, sobre los datos almacenados en registros de tamaño fijo compuestos por campos comparables. Al surgir la posibilidad de almacenar en una base otros tipos de datos tales como los objetos multimediales (imágenes, video, texto) y el hecho de que estos datos no puedan estructurarse, hace necesaria la definicion de nuevas operaciones y capacidad de almacenamiento en las bases de datos. Se espera poder realizar en estos modelos búsquedas eficientemente, teniendo en cuenta cuestiones como que la búsqueda exacta no resulta de interés y que en ciertas ocasiones se requiere mantener los distintos estados de la base de datos a través de tiempo y no sólo el más reciente, a fin de poder realizar consultas de información histórica. Como respuesta a estos requerimientos han surgido modelos como el espacial, temporal, espacio-temporal, espacios métricos y el modelo métrico-temporal, que brindan funcionalidades de persitencia y manipulación de estos tipos de datos. El tema de estudio del *Grupo de Investigación en Bases de Datos (GIBD)*, es el modelado de objetos no estructurados y el procesamiento eficiente de consultas sobre estos tipos de datos.

Palabras Claves: Bases de Datos Espaciales, 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 el ámbito de las bases de datos clásicas se aplica el concepto de búsqueda exacta ya que los datos son estructurados, es decir, que la información se organiza en registros que pueden dividirse en campos cuyos valores son comparables. La respuesta a una búsqueda en este contexto estará conformada por todos aquellos registros cuyos campos coinciden con los ingresados en la consulta. Una característica de las bases de datos clásicas es que en general capturan solamente el estado mas reciente de la realidad modelada, y la evolución de la base de datos se realiza mediante transacciones, descartando el estado previo.

La posibilidad de almacenar y buscar datos no estructurados tales como imágenes, sonido, texto, video, datos geométricos, etc. es una característica novedosa en las bases de datos. Por tal motivo resulta necesario implementar nuevas estrategias de almacenamiento y búsqueda.

En estos nuevos modelos de bases de datos los datos en general no pueden estructurarse, por lo que índices tales como el *B*-Tree* no son utilizables para realizar búsquedas de manera eficiente; otra característica es que en general las consultas por igualdad carecen de interés y, en algunos casos, es un requisito mantener todos los estados de la base de datos y no sólo el más reciente. En este contexto se han generado los nuevos modelos que describimos brevemente a continuación.

Las *Bases de Datos Espaciales* [4] permiten procesar objetos con alguna referencia espacial. 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.

Las *Bases de Datos Temporales* manejan internamente una o más dimensiones temporales, permitiendo asociar tiempos a los datos almacenados. Existen tres clases de bases de datos temporales según el modo 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*.

Los *Espacios Métricos* constituyen un modelo de bases de datos orientado al almacenamiento de objetos no estructurados, 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.

Ante la necesidad de resolver consultas que involucran más de un aspecto de los antes mencionados se plantean combinaciones de los tipos de bases de datos antes mencionado. Así han surgido los modelos *Espacio-Temporal* y *Métrico-Temporal*.

Las *Bases de Datos Espacio-Temporales* tratan con objetos que cambian su identidad, su posición o su forma en el tiempo. 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

Hasta el momento se han propuesto cuatro índices métrico-temporales: el *FHQT-Temporal* [6], el *Historical-FHQT* [2], el *Event-FHQT* [5] y el *Pivot-FHQT* [3] todos ellos han tomado como base el índice para es-

pacios métricos Fixed Height Queries Tree[1], que trabaja con funciones de distancia discretas. Además se han diseñado las variantes *FHQT$^+$-Temporal* y *Event-FHQT$^+$* que permiten tanto funciones discretas como continuas.

Los índices desarrollados hasta el momento han sido evaluados empíricamente con lotes generados a partir de imágenes del sitio *SISAP* (http://www.sisap.org), añadiendo a cada imagen un intervalo de vigencia. Nuestro interés es probar la eficiencia de los índices en aplicaciones reales concretas, para lo cual se ha desarrollado una aplicación que permite realizar consultas métrico-temporales sobre el sistema de archivos de los sistemas operativos (Windows/Linux). Esta aplicación está orientada a la búsqueda por similitud de archivos y carpetas tanto por nombre como por fecha, con diferentes radios de búsqueda, y utiliza índices métrico-temporales que disminuyen significativamente el tiempo de respuesta.

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 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. 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. Se continua trabajando además en el desarrollo de un Sistema de Información Geográfica para el municipio de la localidad de Caseros (Entre Ríos), que permitirá georreferenciar la capa catastral de la localidad y asociar dicha base de datos a la gestión de tasas municipales.

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 y en la actualidad se cuenta con un becario alumno de dicho posgrado, que está comenzando su actividad de investigación. Otra de las integrantes del grupo está también cursando dicha carrera de posgrado. 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. La codirectora del proyecto codirige además una tesis de Maestría en Ingeniería en Sistemas de Información en la temática de bases de datos espacio-temporales.

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] 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.

[6] 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.