Bases de Datos de Objetos No Estructurados

Anabella De Battista , Andrés Pascal,
Pablo Gancharov, Melisa Argüello, Christian Saliwonczyk
Departamento de Sistemas de Información
Fac. Reg. Concepción del Uruguay
Universidad Tecnológica Nacional
Entre Ríos, Argentina
{debattistaa, pascalj, gancharovp, 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

En las bases de datos tradicionales es frecuente el procesamiento de consultas por exactitud o por rango de valores suceptibles de ser ordenados, sobre datos estructurados en registros de tamaño fijo compuestos por campos comparables. La necesidad de almacenar otros tipos de datos tales como los objetos multimediales (imágenes, video, texto) y el hecho de que estos datos no puedan estructurarse, obligó a extender las capacidades de las bases de datos; pero en la mayoría de los casos sólo se permiten el almacenamiento y alguna funcionalidad adicional. Por ello resulta necesario desarrollar nuevos enfoques para almacenar y la buscar objetos no estructurados eficientemente. En estos nuevos modelos la búsqueda exacta carece de interés y en muchos casos se requiere mantener los distintos estados de la base de datos a través de tiempo y no sólo el más reciente, para poder consultar información histórica. Como solución han surgido modelos como el espacial, temporal, espacio-temporal, espacios métricos y el modelo métrico-temporal, que permiten representar y manipular 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 *Métodos de acceso, consultas y aplicaciones en modelos de bases de datos no convencionales* (PID 25-D040) 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

Las bases de datos clásicas se organizan bajo el concepto de búsqueda exacta sobre datos estructurados. Esto significa que la información se organiza en registros los cuales se dividen en campos que contienen valores completamente comparables. Una consulta retorna todos aquellos registros cuyos campos coinciden con los aportados en la consulta (búsqueda exacta). Por otro lado, otra característica importante de las bases de datos clásicas es que capturan sólo un estado de la realidad modelada, usualmente el más reciente. Por medio de las transacciones, la base de datos evoluciona de un estado al siguiente descartando el estado previo.

En la actualidad es necesario implementar nuevas estrategias de almacenamiento y búsqueda para nuevos modelos de bases de datos, que permiten almacenar datos no estructurados tales como imágenes, sonido, texto, video, datos geométricos, etc. Las características principales de estos nuevos tipos de datos es que no poseen una estructura uniforme, por lo cual los índices tales como el *B*-Tree* no se pueden utilizar para hacer más eficiente la búsqueda, 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* 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 vigente (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 Métri-*

co 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 de distancia definida entre los elementos de U que mide 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.

Estos tres tipos de bases de datos se pueden combinar para resolver consultas complejas que involucran más de un aspecto de los anteriormente descriptos. 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, si es que mantienen su dirección. Entre las aplicaciones que tratan con este tipo de bases de datos se incluyen las 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

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

3.1. Consultas Métrico Temporales sobre Cadenas

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 espacios 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.

Para probar la eficiencia de los índices se desarrolló una aplicación que tiene por finalidad permitir efectuar 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.

3.2. Búsqueda de Imágenes

En la búsqueda de imágenes por similitud en grandes bases de datos, es tan importante la eficiencia del sistema (recuperar imágenes en un tiempo razonable) como su eficacia (recuperar imágenes que sean realmente de interés). La eficacia depende principalmente del preprocesamiento de las imágenes, de la técnica de extracción de características y de la función de distancia que se emplee. Por otro lado, los factores de mayor relevancia para la eficiencia del proceso son el costo de la función de distancia y el tipo de índice que se utilice para acelerar la búsqueda.

Un verdadero sistema de recuperación de imágenes debe permitir dar una imagen como objeto de consulta y debe poder determinar la similitud entre ese objeto y cada una de las imágenes de la base de datos en forma eficiente, a fin de responder la consulta.

La búsqueda por similitud aplicada a imágenes implica transformar las imágenes en vectores de características que luego se insertan en un índice métrico. Luego, ante una consulta, se transforma la imagen de consulta de la misma manera para poder buscar usando el índice.

Existen dos tareas que son cruciales en este proceso: una es convertir las imágenes en vectores y la otra definir una función de distancia que permita comparar las imágenes. La primera tarea afecta directamente la eficacia del sistema dado que las búsquedas se realizarán en base a las características extraídas de cada imagen. La segunda tarea afecta tanto la eficacia como la eficiencia; la eficacia porque la función de distancia modela formalmente lo que se entiende por similitud y la eficiencia porque el costo de búsqueda en el índice se ve directamente afectado por el costo de cálculo de la función de distancia y por la distribución de distancias que genera.

Si bien hay numerosos trabajos de investigación que se concentran en el preprocesamiento de las imágenes y extracción de características [8], las funciones de distancia [7] y los índices métricos [4], la mayoría lo hace por separado, sin estudiar la integración de estos aspectos.

En esta línea hemos trabajado definiendo un proceso para el tratamiento integral de las bases de datos de imágenes.

3.3. 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. Se colaboró con el Grupo de Estudios de Calidad y Medio Ambiente de esta Facultad en la elaboración de un informe para analizar y describir el sector comercial y de servicios de la ciudad de Concepción del Uruguay a fin de obtener una herramienta de planificación. Con la Secretaría de Desarrollo Social del Municipio de Concepción del Uruguay se firmó un convenio para desarrollar e implementar una herramienta SIG (Sistema de Información Geográfica) para la LINEA 102 (Línea de los Derechos) que sirva como herramienta de planificación y soporte a la toma de decisiones, mediante la visualización en un mapa de la ciudad de las direcciones asociadas a las denuncias telefónicas recibidas

por dicho servicio de atención telefónica y la vinculación de esta capa con otras de interés. Con la Facultad de Ciencias de la Salud de la Univ. Nac. de Entre Ríos 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. Actualmente se está trabajando 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 y posteriormente servirá como herramienta de planificación para la gestión municipal.

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.

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. 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 tres alumnos becarios que se están formando en estas temáticas y se han desarrollado hasta la fecha cinco tesinas 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] E. Chavez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[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.

[7] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

[8] A. Shahbahrami and D. B. Juurlink. Comparison between color and texture features for image retrieval. In *Proceedings of the 19th Annual Workshop on Circuits, Systems and Signal Processing*, 2008.