Índices para Bases de Datos Métrico-Temporales

Anabella De Battista , Andrés Pascal
Dpto de Sistemas de Información
Universidad Tecnológica Nacional
Fac. Reg. Concepción del Uruguay
Entre Ríos, Argentina
{debattistaa, pascalj}@frcu.edu.ar

Norma Edith Herrera
Departamento de Informática
Universidad Nacional 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

Las bases de datos métrico-temporales constituyen un nuevo modelo de bases de datos orientado al procesamiento de consultas por similitud en un intervalo o instante de tiempo. Este modelo está basado en la combinación de espacios métricos con bases de datos temporales. Para resolver eficientemente consultas métrico-temporales, se han propuesto los índices FHQT-Temporal, que añade intervalos de tiempo a cada nodo de un FHQT, e Historical-FHQT, que consiste en una lista de instantes de tiempo válidos, donde cada elemento contiene el índice FHQT correspondiente a los objetos vigentes en dicho instante. En este artículo se expone este nuevo modelo de base de datos y se describen ambas estructuras de acceso.

1. Introducción

Las operaciones de búsquedas en una base de datos requieren de algún soporte y organización especial a nivel físico. En el caso de las bases de datos clásicas, la organización de la información se basa en el concepto de búsqueda exacta sobre datos estructurados. Esto significa que la información se organiza en registros con campos completamente comparables. Una búsqueda en la base retorna todos aquellos registros cuyos campos coinciden con los aportados en la consulta (búsqueda exacta). Otra característica importante de las bases de datos clásicas es que capturan sólo un estado de la realidad modelizada, usualmente el más reciente. Por medio de las transacciones, la base de datos evoluciona de un estado al siguiente descartando el estado previo.

Actualmente las bases de datos han incluido la capacidad de almacenar otros tipos de datos tales como imágenes, sonido, texto, video, datos geométricos, etc. La problemática de almacenamiento y búsqueda en estos tipos de base de datos difiere notablemente de las bases de datos clásicas en tres aspectos: primero los datos generalmente son no estructurados, esto significa que es imposible organizarlos en registros y campos, segundo la búsqueda exacta carece de interés y tercero resulta de interés mantener todos los estados de la base de datos y no sólo el más reciente.

Es en este contexto donde surgen las bases de datos temporales y los espacios métricos, como nuevos modelos de bases de datos capaces de cubrir eficaz y eficientemente las necesidades de almacenamiento y búsqueda de estas aplicaciones.

Espacios Métricos: este modelo de bases de datos [5] permite trabajar con objetos no estructurados (como imágenes, audio o video) y realizar búsquedas por similitud sobre dichos objetos. Un espacio métrico es un par $(\mathcal{X}, d)$, donde $\mathcal{X}$ es un conjunto de objetos y $d : \mathcal{X} \times \mathcal{X} \rightarrow R^+$ es una función de distancia que modela la similitud entre los elementos de $\mathcal{X}$. La función d cumple con las propiedades características de una función de distancia: $\forall x, y \in \mathcal{X},\ d(x, y) \geq 0$ (positividad), $\forall x, y \in \mathcal{X},\ d(x, y) = d(y, x)$ (simetría), $\forall x, y, z \in \mathcal{X},\ d(x, y) \leq d(x, z) + d(z, y)$(desigualdad tringular). La base de datos es un conjunto finito $\mathcal{U} \subseteq \mathcal{X}$. Una de las consultas típicas que implica recuperar objetos similares de una base de datos es la *búsqueda por rango,* que denotaremos con $(q, r)_d$. Dado un elemento de consulta q, al que llamaremos *query* y un radio de tolerancia r, una búsqueda por rango consiste en recuperar aquellos objetos de la base de datos cuya distancia a q no sea mayor que r. Numerosos índices han sido propuestos para resolver eficientemente este tipo de búsqueda [5, 6, 4, 1]

Bases de Datos Temporales: estas bases de datos incorporan al tiempo como una dimensión, permitiendo almacenar información acerca del pasado, el presente y en algunos casos, pueden predecir el futuro más probable [8]. Existen tres clases de bases de datos temporales según la forma en que manejan el tiempo: *de tiempo transaccional* que registran el tiempo de acuerdo al momento en que se procesan las transacciones, *de tiempo vigente o válido (valid time)*: almacenan el tiempo en que el hecho ocurrió en la realidad, que puede no coincidir con el momento de su registro y *bitemporales* que integran la dimensión transaccional y la dimensión vigente a través del versionado de los estados [10].

Las **Bases de Datos Métrico-Temporales** fueron presentadas en [2, 9, 3] como respuesta a la necesidad de realizar consultas por similitud pero teniendo también en cuenta la dimensión temporal. En este nuevo modelo de bases de datos se permite trabajar con objetos no estructurados con tiempos de vigencia asociados y realizar consultas por similitud teniendo en cuenta el aspecto temporal. Un ejemplo de aplicación de este modelo es el siguiente: en una base de datos donde se registran las huellas digitales de las personas que visitan un museo junto a su fecha y hora de ingreso, ”determinar si una persona estuvo en el museo en una fecha dada (conociendo su huella digital)”. En este trabajo nos proponemos continuar con el estudio del modelo de bases de datos Métrico-Temporales, con el objetivo de diseñar estrategias que permitan mejorar la performance de los índices presentados en [9, 3].

Comenzamos dando una introducción al modelo Métrico-Temporale. Luego presentamos los índices métrico-temporales *FHQT-Temporal* e *Historical-FHQT*, que constituyen las primeras propuestas de estructuras de acceso para este nuevo modelo de basesde datos. Finalizamos explicando el trabajo actual y futuro.

2. El Modelo Métrico-Temporal

Las Bases de Datos Métrico-Temporales permiten realizar búsquedas sobre objetos no estructurados que tienen un intervalo de vigencia asociado y que no poseen un identificador útil como clave de búsqueda, por lo cual tiene sentido realizar consultas por similitud en un instante o intervalo de tiempo.

Sea O el universo de objetos válidos, 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 y desigualdad triangular).

Sea $X \subseteq U$ el conjunto finito sobre el que se realizan las búsquedas, una **consulta por rango métrico-temporal** se denota mediante la 4-upla $(q, r, t_{iq}, t_{fq})_d$ y consiste en recuperar todos los objetos cuyo intervalo de vigencia se superpone en algún punto con el intervalo $[t_{iq}, t_{fq}]$ y que poseen una distancia a q menor o igual que r; en símbolos:

$$(q, r, t_{iq}, t_{fq}) = \{o / (o, t_{io}, t_{fo}) \in X \wedge d(q, o) \leq r \wedge (t_{io} \leq t_{fq}) \wedge (t_{iq} \leq t_{fo})\}$$

En el caso de una consulta instantánea, el tiempo inicial consultado es igual al tiempo final $t_{iq} = t_{fq}$.

Una forma trivial de resolver una consulta métrico-temporal, sin realizar un barrido secuencial sobre todos los elementos de la bases de datos, es construir un índice métrico agregándole a cada objeto el intervalo de tiempo de vigencia del mismo. Luego, ante una consulta $(q, r, t_{iq}, t_{fq})_d$ primero se utiliza el índice métrico para descarta aquellos objetos obj que están a distancia mayor que r de q; posteriormente se realiza un barrido secuencial sobre el conjunto de elementos no descartados por el paso anterior a fin de determinar cuáles objetos son realmente respuesta a la consulta, es decir, cuáles tienen un intervalo de vigencia que se superpone con $[t_{iq}, t_{fq}]$.

La desventaja que tiene esta solución trivial es que no se usa la componente temporal para mejorar el filtrado en el índice; en este proceso sólo se aprovecha la componente métrica. Una mejor estrategia es que durante el proceso de búsqueda se utilice tanto la componente métrica como la componente temporal para descartar elementos.

3. Métodos de Acceso Métrico-Temporales

A continuación presentamos los índices métrico-temporales FHQT-Temporal, e Historical FHQT, que poseen un mejor comportamiento que la solución trivial ante consultas métrico-temporales.

3.1. FHQT-Temporal

Este índice fue presentado en [9] y consiste en una adaptación del índice métrico Fixed Height Queries Tree(FHQT) [1] mediante la incorporación de la dimensión temporal en base a la idea de transformación del índice espacial R-Tree [7] en un índice espacio-temporal.

Un FHQT de k niveles es un árbol que se construye de la siguiente manera: en la raíz se elige un elemento p llamado pivote y para cada distancia $i > 0$ se define el subconjunto de los objetos que se encuentran a distancia i del pivote p. Luego, por cada subconjunto no vacío se crea un hijo de p con rótulo i, y se construye recursivamente un FHQT. Este proceso se repite hasta cubrir los k niveles del árbol, usando siempre el mismo pivote p para todos los nodos del mismo nivel.

El FHQT-Temporal se construye añadiendo un intervalo de tiempo al FHQT en cada uno de sus nodos. Este intervalo representa el período de tiempo máximo de vigencia para todos los objetos del subárbol cuya raíz es dicho nodo. En cada nodo hoja, este intervalo es el período total de vigencia de los objetos que contiene y cada objeto a su vez mantiene su intervalo de vigencia. Para un nodo interior, el intervalo se calcula tomando el tiempo inicial mínimo y el tiempo final máximo de sus hijos.

La figura 1 muestra un ejemplo de un FHQT de dos niveles y su correspondiente versión temporal. En el FHQT (figura izquierda) se ha elegido a u_{10} como pivote del primer nivel y a u_5 como pivote

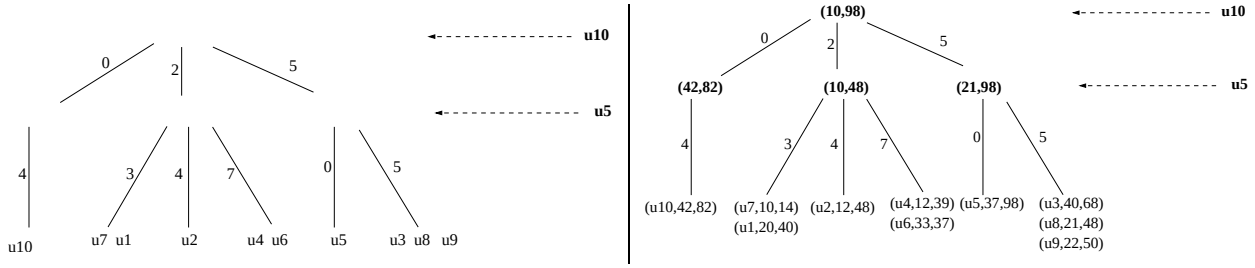

Figura 1: *Un ejemplo de un FHQT (izquierda) y su correspondiente versión temporal.*

del segundo nivel. En este ejemplo u_4 y u_6 se encuentran a distancia 2 del primer pivote y a distancia 7 del segundo pivote. En el FHQT-Temporal (figura derecha) se mantiene la estructura del FHQT incorporando los intervalos de tiempo correspondientes.

Cuando se realiza una consulta métrico-temporal $(q, r, t_{iq}, t_{fq})_d$ se procede de la siguiente manera:

1. En cada nivel del árbol, se descartan todos aquellos subárboles cuyo intervalo de tiempo no tenga superposición con el intervalo $[t_{iq}, t_{fq}]$.

2. Para aquellos subárboles en los que existe superposición temporal, se realiza un filtrado de la misma manera que lo haría una consulta por similitud, es decir, se eliminan todos aquellos subárboles que tengan un rótulo i tal que $i \notin [d(p,q) - r, d(p,q) + r]$ donde p es el pivote del nivel considerado.

3 En el último nivel, se realiza una búsqueda secuencial sobre las hojas que no fueron descartadas, seleccionando aquellos objetos que cumplan con la condición temporal y con la de similitud.

3.2. Historical FHQT

El Historical FHQT (H-FHQT) [3] es otro índice métrico temporal diseñando usando como base el FHQT. Consiste en una lista de los instantes válidos de tiempo, donde cada celda contiene un índice FHQT de todos los objetos vigentes en dicho instante. Esta estructura esta orientada a bases de datos métrico-temporales en donde los objetos tienen vigencia en un sólo instante de tiempo.

La profundidad de el FHQT asociado a cada instante de tiempo varía según la cantidad de elementos que se deban indexar. La cantidad de pivotes a utilizar en un árbol se calcula como $\lceil \log_2 |o_i| \rceil$, donde $|o_i|$ es la cantidad de objetos vigentes en el instante i. Así se evita que haya árboles profundos cuando la cantidad de objetos es baja, para que la estructura no tenga un costo espacial excesivo. La estructura es dinámica, permitiendo altas tanto de objetos en instantes de tiempos ya existentes como de objetos en nuevos instantes.

Las consultas métrico temporales se efectúan de la siguiente manera: en primer lugar se seleccionan los instantes de tiempo i incluidos en el intervalo de consulta. Posteriormente se realizan consultas por similitud usando cada uno de los FHQT correspondientes, y se realiza la unión de los conjuntos resultantes.

.

4. Trabajo Actual y Futuro

Los dos índices presentados en este artículo han demostrado un buen desemepeño en el procesamiento de consultas métrico-temporales [9, 3].

La evaluación experimental del FHQT-Temporal mostró ser más competitivo que la solución trivial, reduciendo significativamente la cantidad de evaluaciones necesarias de la función de distancia, en particular cuando se consulta por instante de tiempo y el radio de búsqueda es relativamente grande. Además permite resolver consultas temporales puras y métricas puras, en este último caso con el mismo costo que el FHQT. Parte de la eficiencia de esta estructura es consecuencia del filtrado inicial por el tiempo, lo que reduce significativamente la cantidad necesaria de evaluaciones de la función de distancia para hallar la respuesta a la consulta.

Actualmente estamos analizando el desempeño del índice en memoria secundaria, tomando la cantidad de accesos a disco como otro factor en el cálculo del costo para esta estructura. Posteriormente pensamos estudiar el problema de la selección de pivotes para la optimización simultánea de la dimensión métrica y temporal.

Por otro lado, el Historical-FHQT ha sido desarrollado como un primer avance de una solución más completa en la que estamos trabajando actualmente. En la misma, introducimos modificaciones al H-FHQT en base a ideas del HR-Tree, para mejorar su eficiencia ante consultas por intervalos, y para permitir representar objetos que tengan asociado un intervalo de vigencia en lugar de un instante.

Referencias

[1] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In *Proc. 5th Combinatorial Pattern Matching (CPM'94)*, LNCS 807, pages 198–212, 1994.

[2] De Battista, A. Pascal, G. Gutierrez, and N. Herrera. Búsqueda en bases de datos métricas-temporales. In *Actas del VIII Workshop de Investigadores en Ciencias de la Computación*, Buenos Aires, Argentina, 2006.

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

[4] E. Chávez, J. Marroquín, and G. Navarro. Fixed queries array: A fast and economical data structure for proximity searching. *Multimedia Tools and Applications (MTAP)*, 14(2):113–135, 2001.

[5] E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[6] Paolo Ciaccia, Marco Patella, Fausto Rabitti, and Pavel Zezula. Indexing metric spaces with m-tree. In *Sistemi Evolui per Basi di Dati*, pages 67–86, 1997.

[7] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *In Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 47–54, 1984.

[8] C. S. Jensen. A consensus glossary of temporal database concepts. *ACM SIGMOD Record*, 23(1):52–54, 1994.

[9] A. Pascal, De Battista, G. Gutierrez, and N. Herrera. Procesamiento de consultas métrico-temporales. In *XXIII Conferencia Latinoamericana de Informática*, pages 133–144, San José de Costa Rica, 2007.

[10] B. Salzberg and V. J. Tsotras. A comparison of access methods for temporal data. *ACM Computing Surveys*, 31(2), 1999.