Búsqueda en Bases de Datos Métricas-Temporales

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

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

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

Resumen

Las bases de datos clásicas se organizan basándose en el concepto de búsqueda exacta sobre datos estructurados, capturando sólo un estado de la realidad modelizada. Actualmente las bases de datos han incluido la capacidad de almacenar otro tipos de datos tales como imágenes, sonido, texto, video, datos geométricos, entre otros. La problemática de almacenamiento y búsquedas sobre estos datos difiere de las bases de datos clásicas y por lo tanto se necesitan nuevos modelos capaces de abordarlos. Entre estos nuevos modelos se encuentran las bases de datos espacios-temporales y los espacios métricos, que han recibido un creciente interés de parte de la comunidad de bases de datos.

En este trabajo nos proponemos estudiar el problema de búsquedas por similitud sobre objetos que pueden variar su identidad, su posición y/o su forma en el tiempo.

Palabras claves: Búsquedas, Bases de Datos Espacio-Temporales, Espacios Métricos,

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 cada uno de los cuales contiene campos completamente comparables. Una búsqueda exacta en la base retorna todos aquellas tuplas cuyos atributos coinciden con los aportados en la consulta.

Una 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 que se presentan en las bases de datos clásicas. En primer lugar los datos generalmente son no estructurados, esto significa que es imposible organizarlos en tuplas compuestas por atributos. En segundo lugar, aún cuando tal estructuración fuera posible, la búsqueda exacta carece de interés en este ámbito; a nadie le interesa, por ejemplo, buscar una imagen que sea exactamente igual a una dada. Y en tercer lugar, en muchas aplicaciones resulta de interés mantener todos los estados de la base de datos y no sólo el más reciente; mas aún, hasta puede surgir la necesidad de predecir un estado futuro de la base.

Es en este contexto donde surgen las bases de datos espacio-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 nuevas aplicaciones.

Las bases de datos espacio-temporales tratan con objetos que cambian su identidad, su posición y/o su forma en el tiempo. Las consultas que se necesitan resolver en este tipo de bases de datos pueden requerir tratar con el tiempo pasado, presente y con predicciones del tiempo futuro. Por ejemplo, nos puede interesar saber cuál es la máxima velocidad alcanzada por un objeto en las dos últimas horas, o encontrar los objetos que cruzaron una cierta área en el tiempo t o en un intervalo de tiempo $[t_1, t_2]$ o buscar objetos que intersectarán un área en movimiento [7, 8, 9, 12].

Otro tipo de búsqueda que resulta de interés en bases de datos no tradicionales es la búsqueda de objetos similares a uno dado . Este tipo de búsqueda tiene una amplia gama de aplicaciones como por ejemplo reconocimiento de imágenes y sonido, compresión de texto, biología computacional, inteligencia artificial y minería de datos, entre otras. Todas estas aplicaciones tienen como características comunes la existencia de un universo de objetos X y de una función de distancia d que modela la similitud entre los objetos del universo. Esto ha dado origen a un nuevo modelo de bases de datos denominado *espacio métrico* [1, 2, 3, 4, 5, 11].

En este trabajo nos proponemos estudiar el problema de búsquedas por similitud sobre objetos que pueden variar su identidad, su posición y/o su forma en el tiempo. Comenzamos dando una breve in-

troducción a la temática de bases de datos espacio-temporales y a la temática de espacios métricos. Luego describimos las búsquedas sobre bases de datos métricas-temporales. Finalizamos explicando el trabajo futuro.

2. Base de Datos Espacio-Temporales

Los sistemas de bases de datos espacio-temporales integran características de las bases de datos espaciales o multidimensionales, con características de las bases de datos temporales, para permitir de manera eficiente, consultas que involucran ambos aspectos. Una aplicación común soportada por este modelo es la que realiza el seguimiento de objetos en movimientos que reportan su ubicación mediante dispositivos GPS. En otras aplicaciones, en lugar de cambiar de ubicación, los objetos pueden cambiar de forma, e incluso de identidad. Los DBMS tradicionales no tienen incorporadas las dimensiones de tiempo y espacio, por lo cual es difícil especificar consultas que combinan estos aspectos.

Inicialmente se desarrollaron los sistemas de bases de datos espaciales, y los sistemas de bases de datos temporales, por separado. Comenzaremos viendo una introducción a cada una de ellas, para luego describir los sistemas de bases de datos espacio-temporales [12].

2.1. Bases de Datos Temporales

Estas bases de datos soportan algún tipo de dominio de tiempo manejado internamente por el sistema administrador de la base de datos [12]. Existen tres clases de bases de datos temporales, en función de la forma en que manejan el tiempo:

De tiempo transaccional (transaction time): registran el tiempo de acuerdo al momento en que se almacena un hecho, es decir, en el orden en que se procesan las transacciones. Hay que notar, que este registro no necesariamente coincide con el orden real en que se produjeron los eventos. Más bien, es acorde al tiempo en que la base tomó conocimiento del evento. Debido a que se mantiene la historia de todos los estados consistentes de la base de datos, se puede realizar un *roll-back* hacia cualquiera de estos estados anteriores. Las bases de datos de tiempo transaccional no permiten modificar el pasado.

De tiempo vigente o válido (valid time): soportan el tiempo en que el hecho ocurrió en la realidad, que puede no coincidir con el momento de su registro. El orden de ocurrencia de los eventos puede diferir del orden de su registro. Este sistema permite realizar correcciones sobre los datos registrados, es decir que los estados anteriores se pueden modificar. En dicho caso, solo se mantiene la última versión de cada estado.

Bitemporales: integran la dimensión transaccional y la dimensión vigente, a través del versionado de los estados, es decir, cada estado se puede modificar para actualizar el conocimiento de la realidad pasada, presente o futura, pero esas modificaciones se realizan generando nuevas versiones de los mismos estados.

2.2. Bases de datos espaciales

Las bases de datos espaciales o multidimensionales ofrecen tipos de datos espaciales en su modelo de datos y un lenguaje de consulta para manipularlos [6]. En un sistema informático estos datos espaciales se representan por puntos, líneas, polígonos, regiones, etc., que se les conoce con el nombre de objetos espaciales. Para responder a consultas relacionadas con propiedades espaciales, se implementan algoritmos eficientes sobre índices espaciales creados a partir de esos objetos.

2.3. Bases de datos espacio-temporales

Los sistemas de Bases de Datos Espacio-Temporales mantienen datos sobre el pasado y el presente y pueden, en algunos casos, realizar predicciones sobre el futuro [10]. Las consultas típicas son de dos clases: *time slice queries* y *time interval (o windows) queries*. Las primeras consultas se realizan sobre un momento dado, como por ejemplo ”buscar todos los objetos que estén en un área en un instante determinado”, mientras que las segundas consultan un intervalo de tiempo, ”buscar todos los objetos que crucen un área entre el momento t_1 y el momento t_2”. Algunas consultas sólo tienen sentido sobre el pasado, otras sólo sobre el presente, otras sobre el futuro, y otras sobre cualquiera de los tres.

Para realizar consultas con eficiencia, se han propuesto varios métodos de acceso que mantienen índices optimizados para tal fin.

Los métodos de acceso en bases de datos espacio temporales se pueden clasificar en tres grupos, de acuerdo al tipo de consulta hacia el cual están orientados:

Recuperación de información histórica: estos métodos permiten responder a las consultas time slice query e interval query, sobre el pasado.

Recuperación de trayectoria: en este caso se quiere mantener la trayectoria que siguen objetos en movimiento.

Predicción de localización: estos métodos permiten calcular la posición futura de los objetos, en base a su posición actual y su patrón de movimiento.

Otra clasificación de los métodos de acceso es según la estrategia de implementación que utilizan:

- Métodos que tratan el tiempo como otra dimensión
- Métodos que incorporan información sobre el tiempo en el índice
- Métodos que utilizan superposición (overlapping) de la estructura, para representar la secuencia de estados en función del tiempo

3. Espacios Métricos

Todas las aplicaciones de búsquedas por similitud nombradas en la sección 1, comparten un marco de trabajo común, que es el de buscar objetos que sean similares bajo alguna función de distancia o similitud adecuada. En esta sección introducimos el modelo formal que abarca todos estos casos.

Denotaremos con $\mathcal{X}$ el universo de objetos válidos. Un subconjunto finito $\mathcal{U} \subseteq \mathcal{X}$, de tamaño n, será el conjunto sobre el que realizaremos las búsquedas. La función:

$$d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$$

denotará una medida de distancia entre objetos de $\mathcal{X}$, esto significa que a menor distancia más cercanos o similares son los objetos. Esta función d cumple con las propiedades características de una función de distancia:

(a) $\forall x, y \in \mathcal{X},\ d(x, y) \geq 0$ (positividad)

(b) $\forall x, y \in \mathcal{X},\ d(x, y) = d(y, x)$ (simetría)

(c) $\forall x, y, z \in \mathcal{X},\ d(x, y) \leq d(x, z) + d(z, y)$ (desigualdad triangular)

El par $(\mathcal{X}, d)$ se denomina *espacio métrico* [5]. La base de datos será un subconjunto finito $\mathcal{U} \subseteq \mathcal{X}$ de cardinalidad n.

En este nuevo modelo de bases de datos, una de las consultas típicas que implica recuperar objetos similares es la *búsqueda por rango*, que denotaremos con $(q, r)_d$. Dado un elemento $q \in \mathcal{X}$, al que llamaremos *query* y un radio de tolerancia r, una búsqueda por rango consiste en recuperar los objetos de la base de datos cuya distancia a q no sea mayor que r, es decir:

$$(q, r)_d = \{u \in \mathcal{U} : d(q, u) \leq r\}$$

Una búsqueda por rango puede ser resuelta con $O(n)$ evaluaciones de distancias examinando exhaustivamente la base de datos. Para evitar esta situación, se preprocesa la base de datos por medio de un *algoritmo de indexación* con el objetivo de construir una *estructura de datos o índice*, diseñada para ahorrar cálculos en el momento de resolver una búsqueda. Un algoritmo de indexación se considera eficiente si puede responder una búsqueda por similitud haciendo una cantidad pequeña de cálculos de distancia, sublineal en la cantidad de elementos de la base de datos.

Básicamente existen dos enfoques para la construcción de algoritmos de indexación en espacios métricos:

Algoritmos basados en pivotes. Estos algoritmos, durante la indexación, seleccionan k pivotes $\{p_1, p_2, \ldots, p_k\}$, y le asignan a cada elemento a de la base de datos, el vector o firma:

$$\Phi(a) = (d(a, p_1), d(a, p_2), \ldots, d(a, p_k))$$

Durante la búsqueda usan la desigualdad triangular junto con la firma de cada elemento para filtrar objetos de la base de datos sin medir su distancia a la query q. Dada $(q, r)_d$, se computa la firma de la query q, $\Phi(q) = (d(q, p_1), d(q, p_2), \ldots, d(q, p_k))$, y luego se descartan todos aquellos elementos a, tales que para algún pivote p_i se cumple que $\mid d(q, p_i) - d(a, p_i) \mid > r$, es decir:

$$max_{1 \leq i \leq k}\{\mid d(a, p_i) - d(q, p_i) \mid\}$$

Los elementos no descartados forman parte de una lista de candidatos, que posteriormente se comparan directamente con la query q.

Algoritmos basados en particiones compactas. Estos algoritmos dividen el espacio en zonas tan compactas como sea posible, y almacenan un elemento (*centro*) representativo de la zona. Junto con el centro se almacena información adicional que permitirá durante la búsqueda descartar aquellas zonas que no contengan elementos de interés.

4. Búsquedas Métricas Temporales

Resulta novedoso realizar búsquedas por similitud pero teniendo en cuenta las componentes espacial y/o temporal. Por ejemplo, supongamos un sistema policial que debe rastrear delincuentes que se movilizan en un automóvil con ciertas características. En este caso, algunas búsquedas que tienen significado son:

- determinar todos los autos similares a uno dado que pasaron por algún lugar en particular.
- determinar todos los autos similares a uno dado en un instante de tiempo.
- determinar todos los autos similares a uno dado que pasaron por algún lugar en particular en un instante de tiempo dado.
- determinar todos los autos similares a uno dado que pasaron por algún lugar en particular en un intervalo de tiempo dado.

Ninguna de estas búsquedas puede ser resuelta eficientemente con índices espacios-temporales dado que los mismos están organizados para realizar búsquedas *exactas* sobre los objetos.

Tampoco podríamos resolverlas con un índice métrico dado que, si bien estos índices permiten realizar búsquedas por similitud, sólo capturan un instante determinado de tiempo y no reflejan la posición del objeto en un espacio físico. Son capaces de determinar la posición o distancia de un elemento respecto de otro, que no es necesariamente una posición física. Por ejemplo: puedo saber cuán lejos o cerca está una imagen de otra dada pero esta distancia está determinada por la similitud de esas imágenes y no por la ubicación en un espacio físico.

En consecuencia, se hacen necesarios nuevos enfoques para atacar esta problemática.

5. Trabajo Futuro

Nos proponemos estudiar las búsquedas métricas-temporales con el fin de proponer una forma eficiente de resolverlas, ya sea proponiendo modificaciones a los índices existentes o diseñando nuevos métodos de acceso ad-hoc.

Las estrategias que se investigarán como parte de este trabajo se pueden resumir en los siguientes puntos:

- Adaptar índices métricos agregando a los objetos el tiempo como un atributo adicional de cada elemento.
- Adaptar índices temporales a fin de que permitan buscar eficientemente por similitud.
- Adaptar índices espaciales a fin de que permitan buscar eficientemente por similitud.
- Adaptar índices espacio-temporales a fin de que permitan buscar eficientemente por similitud.

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] S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.

[3] 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.

[4] E. Chávez and G. Navarro. An effective clustering algorithm to index high dimensional metric spaces. In *Proc. 7th International Symposium on String Processing and Information Retrieval (SPIRE'00)*, pages 75–86. IEEE CS Press, 2000.

[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] Volker Gaede and Oliver Günther. Multidimensional access methods. *ACM Computing Surveys*, 30(2):170–231, 1998.

[7] G. Gutiérrez, G.Ñavarro, A. Rodríguez, A. González, and J. Orellana. A spatio-temporal access method based on snapshots and events. In *Proceedings of the 13th ACM International Symposium on Advances in Geographic Information Systems (GIS'05)*. ACM Press, 2005.

[8] Gilberto Gutiérrez, Gonzalo Navarro, and Andrea Rodríguez. Sest_L: An event-oriented spatio-temporal access method. Technical Report TR/DCC-2006-5, Department of Computer Science, Universidad de Chile (Chile), 2006.

[9] Marios Hadjieleftheriou, George Kollios, Petko Bakalov, and Vassilis J. Tsotras. Complex spatio-temporal pattern queries. In *VLDB*, pages 877–888, 2005.

[10] Mohamed F. Mokbel, Thanaa M. Ghanem, and Walid G. Aref. Spatio-temporal access methods. *IEEE Data Engineering Bulletin*, 26(2):40–49, 2003.

[11] Gonzalo Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.

[12] Betty Salzberg and Vassilis J. Tsotras. Comparison of access methods for time-evolving data. *ACM Comput. Surv.*, 31(2):158–221, 1999.