MÉTODOS DE ACCESO MÉTRICO-ESPACIALES

Andrés Pascal, Anabella De Battista
Grupo de Investigación en Bases de Datos (GIBD)
Facultad Regional Concepción del Uruguay
Universidad Tecnológica Nacional
andrespascal2003@yahoo.com.ar, anadebattista@gmail.com

Norma Herrera
Departamento de Informática
Universidad Nacional de San Luis
nherrera@unsl.edu.ar

RESUMEN

Tradicionalmente, los datos que contienen las bases de datos están estructurados en tuplas y son comparables a través de operadores relacionales. Para acelerar este tipo de consultas existen índices eficientes, tales como B+-Tree. Sin embargo, cada vez es más importante el almacenamiento de objetos no estructurados, que no se pueden comparar por igualdad, para los cuales dichos índices no son aplicables. Algunos ejemplos son: imágenes (rostros, radiografías, pinturas, marcas, paisajes, etc.), texto plano y semiestructurado (documentos, archivos XML, etc.), sonidos (música, voz, etc.) y objetos espaciales (ciudades, rutas, puntos de interés, etc.). Ante esta situación, han surgido otras formas de consultas, siendo algunas de las más importantes las espaciales y las por similitud.

Un aspecto no estudiado aún, es la combinación de estos dos tipos de búsqueda, e.g. "encontrar objetos similares a uno dado, ubicados dentro de un área". Estos tipos de consultas son importantes en especial en los Sistemas de Información Geográfica y aún no existen métodos de acceso que los soporten.

En este proyecto estudiamos distintos aspectos referidos al procesamiento de consultas métrico-espaciales, las funciones de distancia a utilizar, y el uso de paralelismo en GPU para hacer más eficiente el procesamiento de las mismas.

Palabras clave: consultas por similitud, espacios métricos, consultas espaciales, índices

CONTEXTO

Este trabajo se inscribe dentro del Proyecto homologado "Consultas por Similitud y Espaciales en Bases de Datos de Objetos No Estructurados" (UTI3842TC) cuyo objetivo es el desarrollo de métodos y técnicas que mejoren la eficiencia y efectividad de los métodos de búsqueda de objetos no estructurados.

Como parte del proyecto se firmó un convenio con la Municipalidad de Urdinarrain para realizar el relevamiento y geocodificación de los Comercios, Industrias y Estudios Profesionales de dicha ciudad y para el desarrollo de una aplicación que los registre y permita consultas para la toma de decisiones.

Se establecieron comunicaciones con las Oficinas de Marcas y Señales de las provincias de Entre Ríos y Buenos Aires para el desarrollo de una aplicación de búsqueda por similitud de Marcas de Ganado.

Este proyecto es continuación de los proyectos "Procesamiento eficiente de consultas en nuevos modelos de bases de datos" (Incentivos: 25-D059) y "Métodos de Acceso, Consultas y Aplicaciones en Modelos de Bases de Datos no Convencionales" (Incentivos: 25-D040).

1. INTRODUCCIÓN

En las grandes bases de datos, para realizar búsquedas con eficiencia se requiere de algún soporte y organización especial a nivel físico. Las bases de datos clásicas y en particular las relacionales, organizan los datos en conjuntos de registros de tamaño fijo que contienen campos completamente comparables. Esto les permite realizar consultas exactas o por prefijo con costo menor a O(n), mediante estructuras de datos auxiliares llamadas índices, tales como el

B+-Tree o las Hash-tables. Esta manera de organizar los datos y consultar no es adecuada cuando los objetos en cuestión son no estructurados, para los cuales no tiene sentido las comparaciones por igualdad, y donde las consultas incorporan otras dimensiones como el espacio o el tiempo. Por ejemplo, carece de sentido buscar un rostro exactamente igual, píxel a píxel, a algún otro contenido en una base de datos de rostros de personas, o consultar por igualdad una polilínea que representa una ruta dentro de una base de datos espacial.

En este contexto, surgen como respuesta al requerimiento de almacenar y consultar objetos no estructurados y con algún aspecto espacial o temporal, los modelos de Bases de Datos Espaciales, Temporales, Espacio-Temporales, los Espacios Métricos y las Bases de Datos Métrico-Temporales.

Las Bases de Datos Espaciales [1, 2, 3, 4, 5, 6, 7] permiten procesar objetos con alguna referencia espacial y que poseen normalmente una estructura compleja. Un dato espacial se representa usualmente a través de un punto, una polilínea o un polígono. La recuperación y actualización de estos tipos de datos espaciales se basan no sólo en el valor de ciertos atributos, sino también en la ubicación espacial del objeto. Por ejemplo, nos podría interesar obtener los terrenos geográficamente adyacentes a uno dado, o encontrar todas las estaciones de servicio al lado de una ruta. Las bases de datos espaciales se utilizan en muchas áreas, destacándose en particular los Sistemas de Información Geográfica (SIG) [8, 9, 10]. Un SIG es principalmente una herramienta que permite capturar, almacenar, manipular, analizar y mostrar información geográficamente referenciada con el objetivo de resolver problemas complejos de planificación y gestión. En una sociedad donde la información y la tecnología son dos pilares fundamentales, los SIG proveen el marco tecnológico adecuado para el manejo de información geográfica y permiten canalizar la gestión de todo aquello que presente una componente geográfica susceptible de ser aprovechada.

Por otro lado, cuando se almacenan objetos complejos con estructuras variables tales como las imágenes, las consultas que tienen sentido son las llamadas Consultas por Similitud [12, 13, 14]. Una forma de modelar este tipo de bases de datos es mediante Espacios Métricos [11]. Un espacio métrico es un par *(U, d)* donde U es un universo de objetos y *d* es la función de distancia definida entre los elementos de *U*, que mide el grado de similitud (disimilitud, estrictamente hablando) entre ellos y que posee ciertas propiedades que la hacen "métrica".

Una de las consultas típicas en este modelo de bases de datos es la búsqueda por rango: dado un elemento *q* incluido en *U* al que llamaremos *query* y un radio de tolerancia *r*, una búsqueda por rango consiste en recuperar los objetos de la base de datos que se encuentren como máximo a distancia *r* de *q*. Para resolver estas consultas sin realizar O(n) evaluaciones de distancias se utilizan índices que permiten ahorrar cálculos durante el proceso de búsqueda.

En algunos casos, cada vez más frecuentes, las funciones de distancia no cumplen alguna de las propiedades de una función métrica [15, 16]. En especial si la función no respeta la desigualdad triangular, los índices métricos no pueden ser utilizados.

Estos temas constituyen un área de investigación abierta y de gran importancia dado que se basan en necesidades de aplicación reales que aún no están resueltas. Si bien algunos motores de bases de datos como Postgres, Oracle o Informix, ya han incorporado el aspecto temporal o espacial, aún no permiten manejar o tienen limitaciones importantes en cuanto a los demás modelos.

La paralelización de problemas de recuperación de información sobre arquitecturas multi-core, ha sido estudiada desde diversas perspectivas [17, 18, 19, 20, 21, 22, 23]. Por ejemplo, el algoritmo de consulta NN_k (k-Nearest Neighbors) que se utiliza en muchos sistema de recuperación de datos ya ha sido paralelizado con buenos resultados [24, 25].

Actualmente una comunidad importante de investigadores se encuentra dedicada al estudio de la aceleración de diversos algoritmos a través del uso de Unidades de Procesamiento Gráfico (GPU: Graphics Processing Unit). Estas soluciones explotan la naturaleza de las GPUs para alcanzar resultados significativamente más eficientes que los obtenidos mediante el uso

normal de una CPU, utilizando hardware de bajo costo y amplia disponibilidad.
En este proyecto profundizamos el estudio de los distintos aspectos involucrados al procesamiento de consultas por similitud y espacial al mismo tiempo, para las cuales no existen métodos de acceso específicos. En segundo lugar, estudiamos las búsquedas por similitud cuando las funciones no son métricas, incluyendo el desarrollo de estructuras de acceso para estos casos. Y por último, el uso de GPUs para acelerar los mecanismos de búsqueda a través del uso de paralelismo.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Actualmente las líneas de investigación en las cuales se estás desarrollando las actividades del proyecto son:

- Métodos de Acceso Métrico-Espaciales
- Funciones de Distancia y Extracción de Características de Objetos no Estructurados
- Paralelismo sobre GPU

3. RESULTADOS OBTENIDOS/ESPERADOS

En proyectos anteriores se definió el modelo métrico-temporal, se desarrollaron métodos de acceso para procesar consultas que combinan similitud y tiempo, se definieron tipos de consultas sobre este modelo y se construyeron aplicaciones que las soportan.
El objetivo principal de este proyecto es el estudio de nuevas tecnologías para la resolución de Consultas por Similitud y Espaciales, incluyendo:
- Definición del modelo métrico-espacial, incluyendo consultas, métodos de acceso y aplicaciones para este nuevo modelo.
- Estudio y desarrollo de funciones de distancia no-métricas
- Diseño de Métodos de Acceso no-métricos
- Uso de paralelismo (GPUs) para mejorar la eficiencia del procesamiento de consultas métrico-temporales y no-métricas.

Actualmente se han desarrollado funciones de distancia métricas y no-métricas para medir la similitud de imágenes por forma [26, 27], y se ha paralelizado la extracción de características utilizando GPU-Cuda, obteniendo buenos resultados.
También se ha diseñado un primer método de acceso métrico espacial, y se están realizando ajustes y validaciones para realizar los experimentos correspondientes para medir su eficiencia.

4. FORMACIÓN DE RECURSOS HUMANOS

Participan en este proyecto, además de los investigadores a cargo, dos tesistas de la Maestría en Ciencias de la Computación, dos graduados (uno de ellos con beca BINID) de la carrera Ingeniería en Sistemas de Información y todos los años se suman al menos dos alumnos becarios.

5. BIBLIOGRAFÍA

[1] Gandhi, V., Kang, J. M., Shekhar, S.: Spatial Databases, Encyclopedia of Computer Science and Engineering, Wiley, Cassie Craig (Eds.), (2009).

[2] Gaede, V., Günther, O.: Multidimensional access methods. ACM Comput. Surv. 30, 2 (1998).

[3] Baeza-Yates, R., Ribeiro-Neto: Modern Information Retrieval. Addison Wesley (1999).

[4] Beckmann, N., Kriegel, H., Schneider, R., Seeger, B.: The R*-tree: an efficient and robust access method for points and rectangles. ACM International Conference on Magamenet of Data (1990).

[5] Bereczky, N., Duch, A., Németh, K., Roura, S.: Quad-kd trees: A general framework for kd trees and quad trees. Theoretical Computer Science, Volume 616, 22 (2016) ISSN 0304-3975.

[6] Felipe, I., Hristidis, V., Rishe, N.: Keyword search on spatial databases. In ICDE'08, pages 656–665 (2008).

[7] Rouquier, J., Alvarez, I., Reuillon, R., Wuillemin, P.: A kd-tree algorithm to discover the boundary of a black box hypervolume. Annals of Mathematics and Artificial Intelligence (2015).

[8] Chen, Y., Suel, T., Markowetz, A.: Efficient Query Processing in Geographic Web Search Engines. In SIGMOD'06 (2006).

[9] Li, Z., Wang, C., Xie, X., Wang, X., Ma, W.: Indexing implicit locations for geographical information retrieval. In GIR'06 (2006).

[10] Elariss, H., Khaddaj, S.:A time cost optimization for similar scenarios mobile GIS queries. Journal of Visual Languages & Computing (2012).

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

[12] 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).

[13] R. Baeza-Yates. Searching: an algorithmic tour. In A. Kent and J. Williams, editors, Encyclopedia of Computer Science and Technology, volume 37, pages 331-359. Marcel Dekker Inc., (1997).

[14] R. Baeza-Yates and G. Navarro. Fast approximate string matching in a dictionary. In Porc. 5th South American Symposium on String Processing and Information Retrieval (SPIRE'98) (1998).

[15] Scheirer, W., Wilber, M., Eckmann, M., Boult, T.: Good recognition is non-metric. Pattern Recognition, Volume 47 (2014).

[16] Chen, S., Ma, B., Zhang, K.: On the similarity metric and the distance metric. Theoretical Computer Science, Volume 410, Issues 24–25 (2009).

[17] Vinkler, M., Havran, V., Bittner, J.: Performance Comparison of Bounding Volume Hierarchies and Kd-Trees for GPU Ray Tracing. Computer Graphics Forum (2015).

[18] Sun, C., Agrawal, D., Abbadi, A.: Hardware acceleration for spatial selections and joins. In: Proc. ACM Intl. Conf. On Managment of Data. (2003) 455–466.

[19] Bandi, N., Sun, C., Abbadi, A., Agrawal, D.: Hardware acceleration in commercial databases: A case study of spatial operations. In: Proc. Intl. Conf. on Very Large Databases, Morgan Kaufmann (2004) 1021–1032.

[20] Govindaraju, N., Lloyd, B., Wang, W., Lin, M., Manocha, D.: Fast computation of database operations using graphics processors. In: Proc. ACM Intl. Conf. On Managment of Data. (2004) 215–226.

[21] Owens, J., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A., Purcell, T.: A survey of general-purpose computation on graphics hardware. Proc. ACM (2007).

[22] Fatahalian, K., Sugermanm J., Hanrahan, P.: Understanding the Efficiency of GPU Algorithms for Matrix-Matrix Multiplication. Proc. of the ACM SIGGRAPH / EUROGRAPHICS conference on Graphics hardware (2004).

[23] Matsumoto, T., Yiu, M.: Accelerating Exact Similarity Search on CPU-GPU Systems. ICDM, IEEE International Conference on Data Mining (2015).

[24] Cayton, L.: A nearest neighbor data structure for graphics hardware. First International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures (2010).

[25] Bustos, B., Deussen, O., Hiller, S., Keim, D.: A graphics hardware accelerated algorithm for nearest neighbor search. Computational Science ICCS, volume 3994 of LNCS. Springer (2006). PP. 196–199.

[26] G. Sánchez, M. Rodriguez, "Cattle Marks Recognition by Hu and Legendre Invariant Moments". ARPN Journal of Engineering and Applied Sciences, Vol. 11, Nº 1, 2016.

[27] Li S., Lee M., and Pun C., "Complex Zernike Moments Features for Shape-Based Image Retrieval," IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 39, no. 1, pp. 227-237, 2009.