Tecnologías Avanzadas de Bases de Datos

Susana Esquivel, Edilma Gagliardi, Norma Herrera, Verónica Ludueña, Nora Reyes, María Taranilla
{esquivel,oli,nherrera,vlud,nreyes,tarani}@unsl.edu.ar
Departamento de Informática – Universidad Nacional de San Luis
Tel.: 02652-420822-257 – Fax: 02652-430224

Resumen

La evolución de las tecnologías han hecho surgir almacenamientos no estructurados de información. Además de requerir nuevos tipos de datos, en algunos casos no es posible estructurar los datos en forma tradicional mediante claves y registros convencionales. Las nuevas aplicaciones requieren recuperación de información por cualquier atributo, ya sean claves o no, requiriendo hacer uso de herramientas tales como las provistas por la Inteligencia Computacional. Ello ha requerido de nuevos modelos más generales para recuperar y administrar información de este tipo tales como las bases de datos espacio-temporales, de texto, espacios métricos, entre otros.

1. Introducción y motivación

Con la evolución de las tecnologías de información y comunicación, han surgido repositorios o almacenamientos no estructurados de información. No sólo se consultan nuevos tipos de datos tales como datos geométricos, texto libre, imágenes, audio y video, sino que además, en algunos casos, ya no se puede estructurar más la información en claves y registros. Aún cuando sea posible una estructuración clásica, nuevas aplicaciones tales como la minería de datos (data mining) requieren acceder a la base de datos por cualquier campo y no sólo por aquellos marcados como "claves", muchas veces haciendo uso de herramientas no tradicionales, tales como las provistas por la Inteligencia Computacional.

Por lo tanto, se necesitan nuevos modelos para buscar y administrar la información en almacenamientos de este tipo. Los escenarios anteriores requieren modelos más generales tales como las bases de datos espacio-temporales, bases de datos de texto, espacios métricos, entre otros. Se requiere también contar con herramientas que permitan modelar estos tipos de datos, realizar operaciones sobre ellos, definir lenguajes de consulta, etc.

La Geometría Computacional, en Ciencias de la Computación, brinda un marco adecuado para el diseño y análisis de algoritmos para resolver problemas geométricos (y por ende, estructuras de almacenamiento) en un nivel más profundo y conceptualmente mucho más rico [14, 18]. Por lo tanto, una perspectiva geométrica nos permite diseñar y analizar los algoritmos y estructuras de datos utilizadas en bases de datos de tecnología avanzada, con herramientas propias de la Geometría Computacional.

La búsqueda por similitud es un tema de investigación que abstrae varias nociones de las ya mencionadas. Este problema se puede expresar como sigue: dado un conjunto de objetos de naturaleza desconocida, una función de distancia definida entre ellos, que mide cuán diferentes son, y dado otro objeto, llamado la consulta, encontrar todos los elementos del conjunto suficientemente similares a la consulta. El conjunto de objetos junto con la función de distancia se denomina *espacio métrico*. Varios de los problemas que se han mencionado se pueden convertir en problemas de espacios métricos.

En algunas aplicaciones, los espacios métricos resultan ser de un tipo particular llamado "espacio vectorial", donde los elementos consisten de D coordenadas de valores reales. Existen muchos trabajos que explotan las propiedades geométricas sobre espacios vectoriales [9], pero normalmente éstas no se

pueden extender a los espacios métricos generales donde sólo se cuenta con la distancia entre objetos. Se han logrado algunos avances importantes para espacios métricos generales, en su gran mayoría alrededor de la idea de construir un índice, es decir una estructura de datos que reduzca el número de evaluaciones de la distancia durante la consulta [7]. Aunque es muy importante reducir el número de evaluaciones de la distancia, en muchas aplicaciones puede ser también importante reducir además la cantidad de operaciones de E/S realizadas. Algunos trabajos recientes persiguen este doble propósito [8, 15].

2. Geometría Computacional y Bases de Datos

En ocasiones, la Geometría Computacional brinda soluciones más eficientes en problemas que no parecen geométricos. Descubrir que los datos de un problema verifican propiedades geométricas sirve para aplicar alguna técnica algorítmica o alguna estructura de datos especial, que nos permite describir una solución óptima. En nuestra línea, nos dedicamos a diversas temáticas relacionadas a Bases de datos, como Búsquedas por Rangos y Separabilidad Geométrica, Sumas de Minkowski y Algoritmos de Ruteo de Paquetes en Redes Inalámbricas. A continuación brindamos detalles sobre cada temática en particular.

2.1. Las Búsquedas por Rangos

Una de las disciplinas en la que tiene cabida este tema son las bases de datos y el estudio de las consultas por rangos [2, 1]. Aunque, las bases de datos tienen sus aplicaciones más comunes en espacios de datos convencionales, con los avances tecnológicos se han expandido a espacios de datos geométricos.

Un dato geométrico describe un objeto perteneciente al espacio geométrico, que posee características geométricas, como por ejemplo un punto, una recta, un polígono, entre otros. Generalmente van acompañados de datos convencionales y pueden ser discretos o continuos. En el caso discreto (un punto D-dimensional), pueden ser modelados tradicionalmente. Mientras que cuando son continuos (un polígono), cubren una región del espacio y necesitan un tratamiento diferente respecto de las estructuras de datos utilizadas para su almacenamiento y de las técnicas algorítmicas empleadas para su recuperación. La integración de ambos tipos de datos, ha dado lugar a nuevos modelos de bases de datos. Con esto podemos tomar una ligera noción de la necesidad de tener una herramienta teórica de base que nos permita modelar y operar sobre ellos.

Vincular las búsquedas por rangos con separabilidad geométrica nos permite proponer nuevos algoritmos de partición del espacio de búsqueda, por medio de la aplicación de diversos criterios de separabilidad (rectas, cuñas, bandas, entre otras). La aplicación de tales separadores geométricos debe realizarse con relación a la instancia particular del espacio de búsqueda, lo cual hace presuponer, o, conocimientos de tal configuración, o bien, el uso de herramientas no tradicionales que ayuden a la selección de tales criterios de separabilidad, tales como las metaheurísticas [11, 10].

2.2. Sumas de Minkowski

En nuestro trabajo consideramos algunos tipos de operaciones espaciales que específicamente podemos vincular con las sumas de Minkowski, como por ejemplo aquellas que pueden basarse en objetos que “contienen a”, que “son adyacentes a”, entre otras. En este sentido, utilizando la suma de Minkowski, se puede detectar si dos objetos se intersecan (solapados o adyacentes); o también, en el contexto de los sistemas de información geográfica, encontramos consultas referidas a zonas de influencia de ciertas características geográficas, llamadas “zonas buffer”.

Nos hemos dedicado al estudio de su contexto teórico, propiedades geométricas y aplicaciones más destacadas, y además hemos desarrollado una herramienta que implementa la suma de Minkowski entre distintos tipos de polígonos [16, 17].

Inversamente, el problema de descomposición de polígonos resultantes de la suma de Minkowski, es aún un problema abierto. En general nos planteamos si dado un polígono S, existen polígonos P y Q tales que S es la suma de Minkowski de P y Q. Actualmente, nos dedicamos específicamente al problema de la descomposición de polígonos convexos en sumas de Minkowski. El enfoque con el cual estamos trabajando para resolverlo son los Algoritmos Genéticos.

2.3. Algoritmos de Ruteo

Las bases de datos espaciales proporcionan conceptos para bases de datos que siguen la trayectoria de objetos que son móviles en un espacio multidimensional. Si en particular, consideramos una red de computadoras inalámbricas, con nodos móviles, donde en particular interesa el envío de paquetes de un nodo a otro, nos enfrentamos a una temática actual que es el Ruteo de paquetes. Así, nos dedicamos al análisis de algoritmos de ruteo de paquetes en una red de computadoras, tales como las MANets (Movile Ad hoc Networks) y las Redes sin Cables (Wireless Networks), que funcionan sin una infraestructura de conexión fija, donde el ruteo de paquetes debe seguir estrategias diferentes a las conocidas debido a que la organización de la red cambia constantemente.

Hemos fabricado una herramienta que trabaja en la capa de aplicación dentro de un protocolo de redes, obteniendo una colección de métricas que determinan el mejor ruteo de paquetes en cada momento. Las investigaciones realizadas en este ámbito se centran en diversos tópicos, tales como la determinación de las clases de grafos en donde las tasas de éxito en el ruteo de paquetes son elevadas; o bien, en los nodos más utilizados por los algoritmos de ruteo; o sino, en la comparación de los caminos encontrados por estos algoritmos respecto de los óptimos [5, 6, 12]. Se han realizado experiencias cuya finalidad es estudiar la tasa de éxito del ruteo de paquetes y la sobrecarga de nodos en topologías conocidas. Se espera en el futuro ampliar estas investigaciones para el caso de combinaciones de las topologías conocidas, el análisis de las longitudes de caminos y las propuestas de nuevas topologías de red [4].

3. Estudio de índices particionados

Los algoritmos para detección de agrupamientos se han usado en campos del conocimiento humano donde se requiere encontrar una “asociación natural” de algún conjunto de datos específico. Cómo se define esa asociación natural depende del campo y la aplicación particular que emplee esta técnica.

La mayoría de las técnicas se enfocan sobre una función de optimización global. El procedimiento general es proponer un agrupamiento (usando algún algoritmo adecuado) luego medir la calidad y cantidad de los grupos y, si no es satisfactorio, repetir el proceso (por ejemplo con nuevos parámetros) hasta encontrar uno que sí lo sea. Este proceso es suficiente para muchas aplicaciones. Por ejemplo, en Recuperación de la Información se han usado técnicas de detección de agrupamientos tradicionales para propósitos tales como expansión de queries, clasificación de documentos, visualización de resultados, etc.

Aquí nos enfocaremos en un ambiente diferente de detección de agrupamientos. Nuestro objetivo es agrupar datos de un espacio métrico general. Esto implica que la única información de la que disponemos para descubrir la estructura del agrupamiento es la distancia entre elementos del espacio. Mas aún, al no poseer información de coordenadas no contamos con la posibilidad de crear nuevos puntos de datos, tales como centroides. Esto implica que no pueden usarse métodos tradicionales como por ejemplo k *media*, que opera sobre espacios de coordenadas y puede requerir crear nuevos puntos en el espacio. La idea esencial es que, en vez de utilizar una única estructura de datos (índice) para indexar todo el espacio, puede ser conveniente dividir el espacio en dos o más subconjuntos que se indexen y busquen por separado.

Hemos desarrollado un algoritmo de particionamiento de espacios métricos que no se corresponde con la técnica divide y conquistarás. Este algoritmo divide el espacio métrico en dos grupos o núcleos que tienen comportamientos diferentes respecto de su distribución en el espacio. El propósito es mejorar la indexación adecuando los parámetros usados en la construcción del índice a las características de cada uno de los núcleos descubiertos. Es decir, se indexa cada grupo en forma separada adecuando los parámetros del algoritmo de indexación a las propiedades locales de cada grupo. Luego, una búsqueda se resuelve buscando separadamente en cada uno de ellos. Los resultados obtenidos con este algoritmo son alentadores [3], y es por esta razón que continuaremos esta línea de investigación buscando diseñar nuevos algoritmos de detección de agrupamientos.

4. Búsquedas en Bases de Datos no Convencionales

En general para administrar una base de datos conteniendo tipos de datos no convencionales es necesario analizar algunos aspectos teóricos, prácticos y aplicativos del problema. Esto incluye analizar distintos tipos de bases de datos no convencionales, los operadores para responder consultas de sobre ellas, como así también las estructuras y operaciones necesarias para responderlas eficientemente.

Actualmente estamos dedicados a cómo resolver resolver eficientemente las búsquedas por similaridad en distintos tipos de bases de datos no convencionales: conteniendo imágenes, videos, documentos, texto no estructurado; contando con una función de distancia. La necesidad de una respuesta rápida y adecuada, y un eficiente uso de memoria, hace necesaria la existencia de estructuras de datos especializadas que incluyan estos aspectos. Existen índices que, en principio, no sólo resuelven ambos tipos de problemas; pero aún están muy inmaduros para ser usados en la vida real por dos motivos importantes: *falta de dinamismo* y *necesidad de trabajar en memoria principal*. Estas características son sobreentendidas en los índices para bases de datos tradicionales, y la investigación apunta a poner los índices para estas nuevas bases de datos a un nivel de madurez similar.

Integrar la búsqueda de espacios métricos en un ambiente de bases de datos requiere además extender apropiadamente el álgebra relacional y diseñar soluciones eficientes para los nuevos operadores, teniendo en cuenta aspectos de memoria secundaria, concurrencia, confiabilidad, etc., pero además es necesario contar con un índice que permita realizar búsquedas eficientes, que sea dinámico y que además sea competitivo en memoria secundaria. Hemos desarrollado una estructura para búsqueda por similitud en espacios métricos llamado *Árbol de Aproximación Espacial Dinámico (SATD)* [13] que permite realizar inserciones y eliminaciones, manteniendo una buen desempeño en las búsquedas. Muy pocos índices para espacios métricos son completamente dinámicos. Estamos actualmente desarrollando una versión del *SATD* que funcione adecuadamente en memoria secundaria, porque en un escenario real de un ambiente de base de datos es necesario contar con herramientas que permitan trabajar con grandes volúmenes de datos, manteniendo el dinamismo.

Referencias

[1] P. Agarwal. Range searching. In J. Goodmand and J. O'Rourke, editors, *Handbook of Computational Geometry*. CRC Press, 1997.

[2] P. Agarwal and J. Erickson. Geometric range searching and its relatives. In B. Chazelle, J. Goodman, and R. Pollack, editors, *Advances in Discrete and Computational Geometry*. American Mathematical Society, 1998.

[3] R. Baeza-Yates, B. Bustos, E. Chávez, N. Herrera, and G. Navarro. *Clustering in Metric Spaces and Its Application to Information Retrieval*. Kluwer Academic Publishers, 2003.

[4] M. Beron, O. Gagliardi, and G. Hernández. Evaluación de métricas en redes de computadoras. In *Actas del IX Congreso Argentino de Ciencias de la Computación (CACIC'03)*, 2003. Publicación en CD-rom.

[5] P. Bose, A. Brodnik, S. Carlsson, E. Demaine, R. Fleischer, A. López-Ortiz, P. Morin, and J. Munro. Online routing in convex subdivision. In *11th International Symposium on Algorithms and Computation (ISAAC 2000)*, 2000.

[6] P. Bose and P. Morin. Online routing in triangulations. In *10th International Symposium on Algorithms and Computation (ISAAC'99)*, LNCS 1741, pages 113–122. Springer-Verlag, 1999.

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

[8] P. Ciaccia, M. Patella, and P. Zezula. M-tree: an efficient access method for similarity search in metric spaces. In *Proc. of the 23rd Conference on Very Large Databases (VLDB'97)*, pages 426–435, 1997.

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

[10] O. Gagliardi and G. Hernández. Un enfoque propuesto para las búsquedas por rangos con separabilidad geométrica. In *Actas del VIII Congreso Argentino de Ciencias de la Computación (CACIC'03)*, 2003. Publicación en CD-rom.

[11] O. Gagliardi and G.Hernández Peñalver. Las búsquedas por rangos en espacios de búsquedas geométricamente separables. In *Proc. II Workshop de Bases de Datos (JCCC 2003)*, Chillán, Chile, 2003.

[12] E. Kranakis, H. Singh, and J. Urrutia. Compass routing on geometric network. In *Proc. of 11th Canadian Conference on Computational Geometry (CCCG'99)*, 1999.

[13] G. Navarro and N. Reyes. Fully dynamic spatial approximation trees. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002)*, LNCS 2476, pages 254–270. Springer, 2002.

[14] M. Abellanas Oar. Descubriendo la geometría algorítmica. http://www.dma.fi.upm.es/mabellanas/divulgación/GeometriaAlgoritmica.html, 2000.

[15] S. Prabhakar, D. Agrawal, and A. El Abbadi. Efficient disk allocation for fast similarity searching. In *ACM Symposium on Parallel Algorithms and Architectures*, pages 78–87, 1998.

[16] M. T. Taranilla and G. Hernández Peñalver. Descomposición en sumas de minkowski. In *Proc. V Workshop de Investigadores en Ciencias de la Computación (WICC'03)*, 2001.

[17] M. T. Taranilla, M. Printista, and O. Gagliardi. Una propuesta para mejorar el calculo de sumas de minkowski entre polígonos. In *Actas del IX Congreso Argentino de Ciencias de la Computación (CACIC'03)*, 2003. Publicación en CD-rom.

[18] G.T. Toussaint. *Computational Geometry*. North-Holland, Amsterdam, 1985.