Aportes a Nuevos Modelos de Bases de Datos

Jorge Arroyuelo, Maria E. Di Genaro, Alejandro Grosso, Verónica Ludueña, Nora Reyes
Dpto. de Informática, Fac. de Cs. Físico-Matemáticas y Naturales, Universidad Nacional de San Luis
{*bjarroyu, digeme, agrosso, vlud, nreyes*}@unsl.edu.ar

Edgar Chávez
Centro de Investigación Científica y de Educación Superior de Ensenada, México
elchavez@cicese.mx

Rodrigo Paredes
Dpto. de Cs. de la Computación, Fac. de Ingeniería, Universidad de Talca, Chile
raparede@utalca.cl

Resumen

La evolución de las tecnologías de información y comunicación, junto con la gran cantidad y variedad de información disponible digitalmente han llevado al surgimiento de nuevos depósitos no estructurados de información, para los datos que no se adaptan fácilmente al modelo relacional. Estos datos surgen desde campos muy disímiles provocando requerimientos de usuarios que pueden ser tan dispares como el tipo de datos que se necesita procesar. Por tal motivo, es necesario utilizar estos depósitos especializados y formas más sofisticadas de búsqueda sobre los mismos, ya que las soluciones tradicionales no suelen enfrentar tales requerimientos. Para satisfacer estas demandas se deben desarrollar aplicaciones capaces de manipular eficientemente datos muy diferentes entre sí como: secuencias biológicas, huellas digitales, texto, audio, video, imágenes, etc.

Por otro lado, la gran cantidad de datos que se debe manipular para lograr respuestas adecuadas y eficientes, hace necesario un uso eficaz del espacio disponible, lo que implica que las estructuras utilizadas para acceder a este tipo de base de datos, deben ser *conscientes de la jerarquía de memoria*. Un modelo que se adapta a tales requerimientos, en el cual se puede utilizar métodos de acceso que contemplen estos aspectos, son las *Bases de Datos Métricas*. Esta investigación pretende contribuir a la madurez de este nuevo modelo de bases de datos considerando distintas perspectivas.

Palabras Claves: bases de datos no convencionales, índices, búsquedas por proximidad.

Contexto

El presente trabajo se realizó en el marco del Proyecto Consolidado *Tecnologías Avanzadas de Bases de Datos*, (Cód. 03-2218 y en Programa de Incentivos 22-F814) de la Universidad Nacional de San Luis, en la línea *Bases de Datos no Convencionales*. En colaboración con investigadores del Centro de Investigación Científica y de Educación Superior de Ensenada (México) y de la Universidad de Talca (Chile).

La investigación que se realiza en este ámbito, está enfocada en lograr la consolidación de las bases de datos destinadas a manipular datos no convencionales. Esto incluye además, plantear nuevas arquitecturas del procesador que mejoren a muy bajo nivel los administradores de estos nuevos modelos de bases de datos. Se busca así contribuir a distintos campos de aplicación: sistemas de información geográfica, robótica, visión artificial, diseño asistido por computadora, computación móvil, entre otros.

Introducción

La inserción de la computación en casi todos los ámbitos de la sociedad: productivo, artístico, laboral, recreativo, científico, de la salud, etc., hizo necesaria la evolución de las bases de datos. éstas deben ser capaces de administrar todo tipo de datos y responder consultas sobre los mismos de una manera totalmente diferente a la tradicional, muchas veces más intuitiva.

El modelo de *espacios métricos* resulta adecuado para englobar muchas de las características que comparten todas estas aplicaciones, a pesar de ser tan diversas. Formalmente un espacio métrico consiste de un universo de objetos y una función de distancia, definida entre ellos, que mide cuán diferentes entre sí son los objetos. Este contexto permite abordar aplicaciones tales como el ingresar un trozo de melodía en un buscador y que éste presente melodías similares a dicho trozo, o proveerle una imagen y esperar que éste proporcione imágenes semejantes a la suministrada. Las búsquedas tradicionales (exactas) carecen de sentido en la mayoría de estos casos y sobre estos tipos de datos, resultando más naturales las *búsquedas por similitud*, provistas por este modelo.

Para responder eficientemente a este tipo de búsquedas, sin realizar la examinación exhaustiva del conjunto de datos, se necesitan los llamados *Métodos de Acceso Métricos* (MAMs). Sin embargo, además de poder responder eficientemente las búsquedas, es esencial su actualización y optimización ya que la mayoría de estos métodos no admiten actualizaciones (inserciones/eliminaciones), ni están diseñados para soportar conjuntos masivos de datos y tampoco para resolver operaciones de búsquedas más complejas. Los avances en este campo se ven reflejados en áreas tales como: comparación de huellas digitales, reconocimiento de voz, reconocimiento facial, bases de datos médicas, reconocimiento de imágenes, minería de datos, recuperación de texto, biología computacional, clasificación y aprendizaje automático, etc.

Además, en otra de las áreas exploradas, se pretende caracterizar nuevas arquitecturas que permitan reducir el flujo de bits entre el procesador y la memoria, en relación a la cantidad de datos utilizados por cada programa, para mejorar el desempeño a bajo nivel de los administradores de bases de datos (DBMS).

Líneas de Investigación y Desarrollo

Bases de Datos no Convencionales

En este ámbito de investigación, las bases de datos que administran videos, imágenes, texto libre, secuencias de ADN o de proteínas, audio, etc., llamadas *bases de datos no convencionales*, serán modelizadas utilizando el modelo de espacios métricos. Además, a fin de responder eficientemente consultas por similitud se hará uso de MAMs. Debido a lo costoso que generalmente resultan los cálculos de distancia, el número de cálculos realizados al crear el índice o al realizar búsquedas es considerado como medida general de complejidad. Por ello, se analizan aquellos MAMs que han mostrado buen desempeño en las búsquedas, para optimizarlos, siendo conscientes para ello de la jerarquía de memorias.

En general, un espacio métrico consta de un universo $\mathbb{U}$ y una función de distancia d y dada una base de datos $X \subseteq \mathbb{U}$ y una consulta $q \in \mathbb{U}$, las consultas por similitud más comunes son de dos tipos: por *rango* o de *k-vecinos más cercanos* (k-NN).

Métodos de Acceso Métricos

Muchas veces los índices no caben en memoria principal, ya sea porque organizan una base de datos masiva, o porque los objetos de la misma son demasiado grandes. Entonces, surge la necesidad de diseñar índices especialmente pensados para memoria secundaria. Teniendo esto en consideración, se han diseñado dos nuevos índices basados en la *Lista de Clusters*(*LC*) [5] que son totalmente dinámicos, es decir, admiten inserciones y eliminaciones de elementos y están especialmente diseñados para trabajar sobre grandes volúmenes de datos [7]. La *Lista de Clusters Dinámica* (*DLC*), tiene buen desempeño en espacios de alta dimensión, con buena ocupación de página y operaciones eficientes tanto en cálculos de distancia como en operaciones de I/O. Sin embargo, las búsquedas en ella deben recorrer completamente la lista de centros de los clusters, elevando los costos. El *Conjunto Dinámico de Clusters* (*DSC*), también mantiene los clusters en memoria secundaria, pero organiza los centros de clusters en un *DSAT* en memoria principal, permitiendo que las búsquedas realicen menos cálculos de distancia y accedan a menos páginas/clusters. La información de ese *DSAT* también se aprovecha en las inserciones, mejorando los costos de las operaciones en cálculos de distancia y manteniendo los bajos costos de acceso a disco. Ambos, *DLC* y *DSC*, han demostrado tener una razonable utilización de páginas de disco y son competitivas respecto a las alternativas representativas del estado del arte.

Otro aspecto a considerar en este caso es la calidad de los clusters generados. Por lo tanto, una variante para la *DSC*, que está en etapa de evaluación, es que en lugar de insertar los elementos en el índice a medida que van llegando, se puede demorar la incorporación de cada nuevo elemento a un cluster hasta tener varios elementos y poder determinar así

un mejor agrupamiento de dichos elementos. Esto permite además reducir el costo de construcción del índice, porque se realiza una escritura de un cluster en disco luego de varias inserciones e implícitamente puede mejorar los costos de búsqueda al lograr clusters más compactos y que aseguran una total ocupación de la página del disco, achicando así el tamaño del archivo y, por consiguiente, reduciendo los tiempos de acceso. Esta nueva variante, denominada *BOLDSC* (por sus sigla en inglés "Bu?ered On Line Dynamic Set of Clusters"), se espera guarde las buenas características de dinamismo de *DSC*, mejorando su desempeño tanto en construcción como en búsquedas.

Entre la optimizaciones necesarias para los MAM's también se encuentra el dinamismo. Por esta razón, a partir de uno de los índices de mejor desempeño en espacios de mediana a alta dimensión, el *árbol de Aproximación Espacial* (*SAT*), totalmente estático, se desarrolló el *árbol de Aproximación Espacial Dinámico (DSAT)* [7] que permite realizar inserciones y eliminaciones, conservando un muy buen desempeño en las búsquedas, pero agregando un parámetro a sintonizar. Una variante también estática del *SAT*, el *árbol de Aproximación Espacial Distal* (*DiSAT*) [4], logra optimizar las búsquedas respecto de ambos (*SAT* y *DSAT*) y además no necesita ningún parámetro. Por ello, se ha propuesto una optimización del mismo, la *Foresta de Aproximación Espacial Distal* (*DiSAF*) [2], que es dinámica, para memoria principal y que para lograr mejorar al máximo su desempeño, aplica la técnica de dinamización de Bentley y Saxe al *DiSAT* y aprovecha el profundo conocimiento que se tiene sobre la aproximación espacial. Sin embargo, como en este método cada inserción de un elemento provoca la reconstrucción parcial de la estructura, es costosa de construir. Por ello, se ha diseñado una nueva versión dinámica del *DiSAT*, que demora las reconstrucciones para amortizar su costo entre muchas inserciones. A este nueva forma de inserción se la denomina *inserción perezosa*. La ventaja es que logra bajar los costos de construcción, manteniendo el buen desempeño en las búsquedas.

Otra faceta que hay que tener en cuenta, son los requerimientos de algunas aplicaciones que priorizan la rapidez en las respuestas aunque sea a costa de perder algunos elementos: se intercambia precisión (devolviendo sólo algunos objetos relevantes) por velocidad en la respuesta. Este tipo de búsquedas se denominan *aproximadas*. Para conjuntos de datos masivos, las búsquedas por similitud aproximadas permiten obtener un buen balance entre el costo de las búsquedas y la calidad de la respuesta obtenida. En este contexto, se están analizando distintas posibilidades de utilizar búsquedas aproximadas para acelerar las búsquedas, pero intentando mantener una buena calidad de respuesta.

Problema de los k Vecinos

Entre las búsquedas por similitud más utilizadas, la de los k vecinos más cercanos resulta útil en la predicción de funciones, en el aprendizaje automático, la cuantificación y compresión de imágenes, entre otras. Una generalización de la misma consiste en obtener los k-vecinos más cercanos de *todos* los elementos de la base de datos (*All-k-NN*). El planteo general de esta consulta puede verse como: para cada elemento $u \in X$, donde la base de datos $X \subseteq \mathbb{U}$, se responde con los k elementos en X-$\{u\}$ que tengan la menor distancia d a u. La solución burda a este problema tiene una complejidad de n^2 cálculos de distancia, con $|X| = n$; ésta consiste en comparar cada elemento de la base de datos con todos los demás. Una solución más eficiente implica el preprocesamiento de los datos, por ejemplo por medio de un índice, para reducir el número de cálculos de distancia en las búsquedas.

El *Grafo de los k-vecinos más cercanos* (kNNG) [8] se encuentra entre las soluciones propuestas para espacios métricos generales y su desempeño supera algunas de las técnicas clásicas. La construcción del kNNG permite indexar un espacio métrico y luego él mismo se emplea en la resolución de las consultas por similitud. Sin embargo, existen situaciones en las cuales el costo de la construcción del índice, para luego obtener los vecinos más cercanos, puede resultar excesivo. Por ejemplo, cuando la función de distancia es demasiado costosa de calcular, o si se tiene una base de datos masiva, o cuando se pretende resolver consultas en espacios métricos de alta dimensión, donde muchas veces se requiere revisar casi todo el conjunto de datos sin importar la estrategia utilizada. Otro factor a tener en cuenta son los requerimientos de algunas aplicaciones particulares, que priorizan la velocidad de respuesta sobre la precisión de la misma [9, 5, 10, 6]. Para hacer frente a éstas circunstancias es que se han considerado las llamadas *búsquedas por similitud aproximadas*. Este tipo de consultas aceptan algunos "errores" en la respuesta, si con esto se mejora la complejidad de la misma.

Para tal fin, se han desarrollado algunas propuestas que computan una aproximación del kNNG, es decir conectan cada objeto u de la base de datos con k vecinos *cercanos*, relajando la condición que exige que no haya, en toda la base de datos, algún objeto más cercano a u que los k vecinos devueltos. Esto puede ocasionar que se pierda algún objeto muy cercano y en su lugar se devuelva otro un poco más lejano, pero a cambio la respuesta será más rápida. A este grafo se lo denominó *Grafo de vecinos cercanos* (knNG) [3]. Una característica común que tienen estos desarrollos es que ninguno utiliza un índice para buscar en él. Una primera aproximación aprovecha el profundo conocimiento que se tiene del *DiSAT* para plantear un enfoque novedoso. Aquí se consideró un caso particular del problema, cuando $k = 1$, obteniendo el 1nNG. Esta propuesta utiliza la información obtenida durante la *construcción* del *DiSAT* para construir el 1nNG, conectando a cada elemento u con un elemento cercano de la base de datos, que puede ser, o no, su vecino más cercano [3]. Esta propuesta permite recuperar el 1nNG con bajo costo, una muy buena precisión y un error bajo, logrando un buen compromiso calidad/tiempo, y llamativamente *sin realizar ninguna búsqueda.*

Las otras propuestas abordadas se enfocan en responder a los *All-k*-nN y computar el knNG. Estos planteos no utilizan el apoyo de ningún índice, no sólo no buscan en ellos, sino que ni siquiera recurren a la información provista por su construcción. El propósito de estos desarrollos es aprovechar de manera ingeniosa las propiedades de la *función de distancia*. En ellos se sugieren distintas maneras de seleccionar muestras de la base de datos, a partir de las cuales se obtiene un conjunto de distancias que serán el punto de partida de este proceso. También se analizan diferentes maneras de utilizar la información conseguida, para calcular los vecinos aproximados para todos los objetos de la base de datos, utilizando propiedades como la simetría o la desigualdad triangular. Los resultados de estas propuesta se muestran muy prometedores.

Arquitecturas de Procesadores Orientadas a Bases de Datos

Hay autores que realizan una distinción entre arquitectura, implementación y realización. La arquitectura de una computadora define el conjunto mínimo de propiedades que determinan qué programas correrán y qué resultados producirán sobre el procesador, es decir es la interfaz entre el software y el hardware de una computadora. La organización básica del flujo de datos y el control, conforma la implementación. La estructura física que comprende la implementación, conforma la realización [1].

En la actualidad, la investigación sobre arquitecturas de procesadores ha sido desplazada por investigación sobre la implementación de procesadores. La mayoría de los trabajos de investigación están dedicados a mejorar técnicas de predicción (tanto de control como de datos), técnicas para sincronizar y comunicar procesadores (núcleos) a través de mensajes y/o memoria compartida. Muchas de estas técnicas de implementación surgieron en los años 60 y hoy se han incorporado a los diseños de microprocesadores actuales. No obstante, estas técnicas de implementación se pueden aplicar a todo tipo de arquitecturas, desde una arquitectura RISC [1] (que intenta acercar el lenguaje de máquina al hardware del procesador) a una arquitectura que se aleje del hardware e intente disminuir el tráfico de bits entre procesador y memoria.

Si bien las arquitecturas RISC compitieron en desempeño con las arquitecturas CISC [2], las mismas poseen un alto tráfico de bits entre el procesador y la memoria para una determinada traza de ejecución. Esto finalmente favoreció a las CISC sobre las RISC, una vez que las CISC mejoraron sus técnicas de implementación. El objetivo de las investigaciones en esta línea de trabajo es plantear nuevas arquitecturas que minimicen el tráfico de bits entre el procesador y la memoria. Para esto se está realizando la construcción de un simulador del set de instrucción AMD-64 o x86-64 con el fín de evaluar el tráfico de bits para benchmarks como Specint y Specfp para la arquitectura x86. El próximo paso sería evaluar el tráfico de bits para la arquitectura propuesta sobre los mismos benchmarcks, lo cual implica construir no sólo el simulador de la arquitectura sino también el compilador C para la misma. Finalmente, se pretende aprovechar el conocimiento adquirido para, desde bajo nivel, mejorar el desempeño de los DBMSs.

Resultados y Objetivos

Los resultados obtenidos en las investigaciones realizadas sobre el modelo de espacios métricos, además de permitir mejorar el desempeño de los MAMs analizados, conducen a estudiar su aplicación a otros métodos de acceso [2, 3, 4, 7].

[1] acrónimo del inglés "Reduced Instruction Set Computer".

[2] Acrónimo del inglés "Complex Instruction Set Computer".

Se ahondará en el estudio de nuevos diseños de estructuras de datos, capaces de adaptarlas tanto al nivel de la jerarquía de memorias donde se almacenarán, como a las características de los datos a ser indexados, mejorando así su eficiencia. Se espera brindar nuevas herramientas de administración para bases de datos métricas eficientes, que permitan una madurez cada vez más cercana a la de los modelos tradicionales de base de datos.

Se profundizará el análisis sobre cómo resolver consultas de manera más eficiente, sin la utilización de índices, y sobre distintos espacios. Además, se espera mejorar el desempeño de las operaciones de bajo nivel en los DBMS, mediante una nueva arquitectura del procesador.

Actividades de Formación

Dentro de esta línea de investigación se forman alumnos y docentes-investigadores en:

- **Doctorado en Cs. de la Computación:** una tesis en desarrollo sobre expresividad de lenguajes lógicos de consulta.
- **Maestría en Cs. de la Computación:** una tesis sobre búsqueda por similitud aproximada, próxima a su defensa, y otra en desarrollo sobre un índice dinámico eficiente.
- **Maestría en Informática**: una tesis finalizada, de la Universidad Nacional de San Juan, sobre un índice dinámico para búsquedas aproximadas, especialmente diseñado para memoria secundaria.

Referencias

[1] G. Blaauw and F. Brooks, Jr. *Computer Architecture: Concepts and Evolution.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1997.

[2] E. Chávez, M. Di Genaro, N. Reyes, and P. Roggero. Decomposability of disat for index dynamization. *Computer Science & Technology*, pages 110–116, 2017.

[3] E. Chávez, V. Ludueña, N. Reyes, and F. Kasián. All near neighbor graph without searching. *Computer Science & Technology*, 18(1):61–67, April 2018.

[4] E. Chávez, V. Ludueña, N. Reyes, and P. Roggero. Faster proximity searching with the distal {SAT}. *Information Systems*, 59:15 – 47, 2016.

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

[6] P. Ciaccia and M. Patella. Approximate and probabilistic methods. *SIGSPATIAL Special*, 2(2):16–19, 2010.

[7] G. Navarro and N. Reyes. New dynamic metric indices for secondary memory. *Information Systems*, 59:48 – 78, 2016.

[8] R. Paredes, E. Chávez, K. Figueroa, and G. Navarro. Practical construction of k-nearest neighbor graphs in metric spaces. In *Proc. 5th Workshop on Efficient and Experimental Algorithms (WEA)*, LNCS 4007, pages 85–97, 2006.

[9] H. Samet. *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

[10] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. XVIII, 220 p., Hardcover ISBN: 0-387-29146-6.