# *Consolidación de un Modelo para Bases de Datos no Convencionales*

Jorge Arroyuelo, Susana Esquivel, Alejandro Grosso, Verónica Ludueña, Nora Reyes
Dpto. de Informática, Fac. de Cs. Físico-Matemáticas y Naturales, Universidad Nacional de San Luis
{*bjarroyu, esquivel, agrosso, vlud, nreyes*}*@unsl.edu.ar*
Edgar Chávez
Escuela de Ciencias Físico–Matemáticas, Universidad Michoacana de San Nicolás de Hidalgo
*elchavez*@umich.mx
Gonzalo Navarro
Departamento de Ciencias de la Computación, Universidad de Chile.
*gnavarro@dcc.uchile.cl*

## Resumen

Por la evolución de las tecnologías de información y comunicación, han surgido en la actualidad aplicaciones no tradicionales sobre bases de datos que contienen datos no estructurados tales como texto libre, imágenes, audio, video, secuencias de ADN, etc., provenientes de diversas fuentes como revistas, transacciones financieras, fotografías, música, etc., Además estos datos *multimedia* no pueden ser consultados de manera significativa en el sentido clásico, todas las consultas son por objetos similares a uno dado. Estos escenarios requieren modelos más generales, como las *Bases de Datos Métricas*, con una madurez semejante al de las bases de datos tradicionales.

Por otro lado, el desarrollo de memorias más rápidas y de gran capacidad, promovió la aparición de estructuras de datos que tienen en cuenta estas arquitecturas como las *estructuras de datos con I/O eficiente*. Además, en algunos casos los lenguajes de consulta poseen poco poder expresivo para poder expresar todas las consultas consideradas de interés en este modelo de base de datos. Nuestra investigación pretende contribuir a la consolidación de este nuevo modelo de bases de datos.

***Palabras Claves:*** bases de datos no convencionales, lenguajes de consulta, índices, expresividad.

## Contexto

En el Proyecto Consolidado 330303 “Tecnologías Avanzadas de Bases de Datos” se encuentra la línea *Bases de Datos no Convencionales*, la cual motiva esta presentación. Este proyecto pertenece a la Univ. Nac. de San Luis y se encuentra dentro del Programa de Incentivos a la Investigación (Código 22/F014). El ámbito de este proyecto ha permitido el estudio y tratamiento de objetos de diversos tipos, útiles en distintos campos de aplicación: sistemas de información geográfica, robótica, visión artificial, computación móvil, diseño asistido por computadora, motores de búsqueda en internet, computación gráfica, entre otras, y que se relacionan en tales bases de datos. Se consideran como actividades centrales de esta línea el análisis de distintos tipos de bases de datos, la investigación de aspectos empíricos, teóricos y aplicativos derivados de la administración de una base de datos que maneja tipos de datos no convencionales, la expresividad de los lenguajes de consulta, los operadores necesarios para responder consultas de interés, y también las estructuras y operaciones necesarias para resolverlas eficientemente.

El contacto permanente con investigadores de otros países permite nuevas perspectivas en nuestras investigaciones, gracias a la participación de nuestros integrantes en proyectos conjuntos de cooperación internacional con: Universidad de Chile, Universidad de Massey (Nueva Zelanda), Universidad Michoacana de San Nicolás de Hidalgo (México).

## Introducción

La evolución de las nuevas tecnologías dio lugar a diversas aplicaciones en las cuales el modelo clásico de bases de datos no puede aplicarse, ya sea por el tipo de datos que ellas manejan, como por las exigencias de las mismas. Este escenario requiere modelos más generales, además la necesidad de una respuesta rápida y adecuada, y un eficiente uso del espacio disponible, hacen necesaria la existencia de estructuras de datos especializadas que incluyan estos aspectos. Todas estas aplicaciones tienen características comunes, capturadas en el modelo de *espacio métrico*. Formalmente, un espacio métrico consiste de un universo de objetos $\mathbb{U}$ y una función de distan-

cia definida entre ellos $d : \mathbb{U} \times \mathbb{U} \longmapsto \mathbb{R}^+$ que mide la (di)similitud entre los objetos. En este ámbito las búsquedas exactas carecen de sentido y es importante la elección de este modelo por las *búsquedas por similitud*, más naturales sobre estos tipos de datos.

El trabajo con bases de datos masivas, o con bases de datos que almacenan objetos muy grandes, da lugar a líneas de investigación que, conscientes de estos problemas de costos, diseñan estructuras de datos más eficientes para memorias jerárquicas.

La conocida "*maldición de la dimensionalidad*", por la cual el desempeño de los índices existentes se deteriora exponencialmente con la dimensión del espacio, afecta tanto a espacios de vectores (representación común de datos multimedia) como a espacios métricos, aunque en éstos últimos no está completamente analizado su efecto sobre los MAMs (*métodos de acceso métricos*). De las numerosas estructuras que existen para búsquedas por similitud en espacios métricos, sólo pocas son eficientes en espacios de alta o mediana dimensión, y la mayoría no admiten dinamismo, ni están diseñadas para conjuntos masivos de datos (en memoria secundaria). Por lo tanto, nos dedicamos a estudiar distintas maneras de optimizarlas.

Además del dinamismo en las estructuras y operaciones de búsqueda complejas, también se están investigando son la obtención de mayor expresividad en los lenguajes utilizados para expresar consultas y caracterizar la clase de consultas computables.

## Líneas de Investigación y Desarrollo

### Bases de Datos Métricas

Tomando como modelo para las bases de datos no convencionales a los espacios métricos, es necesario responder consultas por similitud eficientemente haciendo uso de MAMs. En espacios métricos generales la complejidad usualmente se mide como el número de cálculos de distancias realizados. Por ello, se analizan aquellos MAMs que han mostrado buen desempeño en las búsquedas, para optimizarlos más, considerando la jerarquía de memorias. En general, dada una base de datos $X \subseteq \mathbb{U}$ y un objeto de consulta $q \in \mathbb{U}$ las consultas son de dos tipos: por *rango* o de $k$*-vecinos más cercanos*.

#### Métodos de Acceso Métricos

El estudio del *Árbol de Aproximación Espacial* [12], que había mostrado un muy buen desempeño en espacios de mediana a alta dimensión, pero totalmente estático, nos permitió el desarrollo de un nuevo índice llamado *Árbol de Aproximación Espacial Dinámico (DSAT)* [13] que permite realizar inserciones y eliminaciones, conservando su buen desempeño en las búsquedas, lo cual es importante porque pocos índices son completamente dinámicos.

El *DSAT* es una estructura que particiona el espacio considerando la proximidad espacial; pero, si el árbol agrupara los elementos muy cercanos entre sí, lograría mejorar las búsquedas al evitar recorrerlo para alcanzarlos. Podemos pensar entonces que construimos un *DSAT*, en el que cada nodo representa un grupo de elementos cercanos ("clusters") y los relacionamos por su proximidad en el espacio. Cada nodo mantiene el centro del cluster correspondiente, y almacena los $k$ elementos más cercanos a él; cualquier elemento a mayor distancia del centro que los $k$ almacenados, forma parte de otro nodo en el árbol [2]. Sin embargo, falta analizar cuán bueno es el agrupamiento o "clustering"que logra esta estructura, lo cual se podría analizar con nuevas estrategias de optimización de funciones a través de heurísticas bioinspiradas, que han mostrado ser útiles en detección de clusters.

Al trabajar sobre base de datos métricas, puede surgir la necesidad de hacer uso de la memoria secundaria. Es posible que la base de datos no pueda almacenarse en memoria principal por ser masiva o porque sus objetos son muy grandes, o que el índice no quepa en memoria principal, o ambas cosas. Por lo tanto, existe la necesidad de diseñar los índices especialmente para memoria secundaria. Así, en [14] se presentaron versiones preliminares del *DSAT* (*DSAT+* y *DSAT**) especialmente diseñadas para memoria secundaria: índices con buena ocupación de página y eficientes tanto en el número de cálculos de distancia y de operaciones de I/O para cada operación, y se están analizando variantes que mejoren aún más su desempeño. En numerosas aplicaciones es muy importante mantener el total dinamismo de las estructuras, es decir soportar tanto inserciones como eliminaciones de elementos, por lo tanto otro aspecto sobre el que se está trabajando es lograr que las operaciones de inserción y eliminación de elementos en las versiones de memoria secundaria del *DSAT* sean eficientes.

#### Join Métricos

El modelo de espacios métricos permite cubrir muchos problemas de búsqueda por similitud, aunque en general se deja fuera de consideración al operador de ensamble o "join" por similitud, otra primitiva importante [5].

De hecho, a pesar de la atención que esta primitiva ha recibido en las bases de datos tradicionales y aún en las multidimensionales, no han habido grandes avances para espacios métricos generales. Nos hemos planteado resolver algunas variantes del problema de join por similitud: (1) *join por rango:* dadas dos bases de datos de un espacio métrico y un radio $r$, encontrar todos los pares de objetos (uno desde cada base de datos) a distancia a lo sumo $r$, (2) *$k$–pares más cercanos:* encontrar los $k$ pares de objetos más cercanos entre sí (uno desde cada base de datos). Para resolver estas operaciones de manera eficiente hemos diseñado un nuevo índice métrico, llamado *Lista de Clusters Gemelos* (*LTC*) [18], éste se construye sobre ambas bases de datos conjuntamente, en lugar de indexar una o ambas bases de datos independientemente. y permite también resolver las consultas por similitud clásicas sobre cada una de las bases de datos independientemente.

A pesar de que esta estructura ha mostrado ser competitiva y obtener buen desempeño en relación a las alternativas más comunes para resolver las operaciones de join, queda mucho por mejorar para que se vuelva una estructura práctica y mucho más eficiente para trabajar con grandes bases de datos métricas. A la fecha se está analizando la construcción de otra clase de índice basada en "permutantes" para resolver el join aproximado de dos bases de datos métricas; es decir que permita rápida y eficientemente encontrar los pares de elementos más similares entre ambas bases de datos, aunque no los obtenga a todos. Así sería posible extender apropiadamente el álgebra relacional como lenguaje de consulta y diseñar soluciones eficientes para nuevas operaciones, considerando aspectos de memoria secundaria, de concurrencia, de confiabilidad, etc. Algunos de estos problemas ya poseen solución en bases de datos espaciales, pero no en bases de datos métricas.

**Búsqueda aproximada de los *All-$k$*-NN**

Hay muchas otras aplicaciones, tales como la clasificación y aprendizaje automático, donde un nuevo elemento debe ser clasificado de acuerdo a sus vecinos más cercanos, la cuantificación y compresión de imágenes, donde sólo algunos vectores pueden ser representados y los que no deben ser codificados como su punto representable más cercano, la predicción de funciones, en la que desea buscar el comportamiento más similar de una función en el pasado para predecir su comportamiento futuro probable, etc. Todas estas aplicaciones tienen características comunes y necesitan estructuras de datos especializadas que las incluyan, como las de *espacios métricos*.

Dado que en muchas aplicaciones la evaluación de la función de distancia $d$ suele ser una operación muy costosa, se usa como medida de complejidad en la mayoría de los casos. Las investigaciones en la actualidad tienden al estudio de algoritmos en espacios métricos generales, donde existen varias técnicas conocidas para resolver el problema de consultas por similitud en un número sublineal de cálculos de distancia, con la condición del preprocesamiento del conjunto de datos.

Si consideramos las búsquedas por similitud vemos que la recuperación de los $k-$vecinos más cercanos, es uno de sus primitivos básicos. Este puede definirse como: Sea $X$ un conjunto de elementos y la función de distancia definida entre ellos $d$ los $k$-NN($u$) son los $k$ elementos en $X - \{u\}$ que tengan la menor distancia a $u$ de acuerdo con la función $d$. Una variante de este problema, quizá menos estudiada, es la búsqueda de los $k-$vecinos más cercanos de *todos los elementos* de $X$, *All-$k$*-NN, es decir: Sea $u_i \in X$, obtener los *All-$k$*-NN es calcular los $k$-vecinos más cercanos para *todos* los $u_i$ en $X$, por supuesto realizando menos de $n^2$ cálculos de distancia. Así, en el marco de una etapa de investigación previa, se desarrollaron y propusieron soluciones a este problema [17, 16] basadas en la construcción del *Grafo de los $k$-vecinos más cercanos* ($k$NNG) para indexar un espacio métrico requiriendo una cantidad moderada de memoria y la utilización del mismo en la resolución de las consultas por similitud en espacios métricos generales. El $k$NNG es un grafo dirigido ponderado que conecta cada elemento del espacio métrico mediante un conjunto de arcos cuyos pesos se calculan de acuerdo a la métrica del espacio en cuestión. El desempeño en las búsquedas por similaridad de esta propuesta es superior al obtenido utilizando las técnicas clásicas basadas en pivotes.

Por otro lado, el compromiso de tratar de realizar la menor cantidad de cálculos de distancias posibles durante una búsqueda, ha llevado a investigar un enfoque *aproximado* eficiente para resolver estas consultas por similitud. Este enfoque consiste en permitir una relajación en la precisión de la consulta con el fin de obtener una aceleración en la complejidad del tiempo de consulta [19, 6, 15]. El objetivo de la búsqueda por similitud aproximada es reducir significativamente los tiempos de búsqueda al permitir algunos errores en el resultado de la consulta. Adicionalmente a la consulta se especifica un parámetro

$\varepsilon$ de precisión para controlar cuán lejos queremos el resultado de la consulta del resultado correcto. Un comportamiento razonable para este tipo de algoritmo es acercarse asintóticamente a la respuesta correcta como $\varepsilon$ se acerca a cero. Por lo tanto, el éxito de una técnica de aproximación se basa en la resolución del compromiso calidad/tiempo [3]. Esta alternativa a la búsqueda por similitud "exacta"se llama *búsqueda de similitud aproximada* [3], y abarca algoritmos aproximados y probabilísticos.

### Lenguajes de Consulta

La relación existente entre lógica y teoría de bases de datos es muy estrecha y natural, ya que es posible pensar en una base de datos simplemente como una estructura finita, y utilizar las lógicas para expresar consultas sobre éstas. Esto les da una posición central como modelo computacional para el análisis del poder expresivo de los lenguajes de consultas que nos permiten obtener información de una base de datos, siendo relevante como marco teórico para el estudio de las bases de datos.

La mayoría de los lenguajes de consulta sobre bases de datos es equivalente, en su poder expresivo, a $FO$ (First-Order logic). El principal problema es que la expresividad de $FO$ no es lo suficientemente poderosa, porque no alcanza para reflejar ciertas consultas. Esto ha llevado a la búsqueda de una mayor expresividad por medio de diferentes mecanismos de extensión sobre $FO$ utilizados como herramientas de construcción de lógicas más poderosas. Uno de ellos gracias a incorporar cuantificadores que no pueden ser expresados en $FO$, como *clausura transitiva* y *punto fijo*, entre otros, los que han sido ampliamente estudiados. La idea de agregar cuantificadores es generalizada mediante la noción de *cuantificadores generalizados de Lindström*[8]. Aún así, estas lógicas todavía resultan incompletas, por lo que se analizan lógicas de orden superior, $SO$ (Second-Order Logic), y algunos de sus fragmentos que han demostrado poseer propiedades interesantes sobre las estructuras finitas. Un resultado importante de R. Fagin fue la caracterización del fragmento existencial $SO\exists$[7]. Allí se establece que las propiedades de las estructuras finitas que son definidas por sentencias existenciales de segundo orden coinciden con las propiedades de la clase de complejidad NP, lo cual fue extendido por Stockmeyer [20], estableciendo una relación cercana entre la lógica $SO$ y la jerarquía de tiempo polinomial ($PH$).

Actualmente existen muchos resultados igualando la expresividad lógica a la complejidad computacional, pero requieren estructuras ordenadas [9], [10]. Estas relaciones entre la complejidad computacional (cantidad de recursos necesarios para resolver un problema sobre algún modelo de máquina computacional) y la complejidad descriptiva (el orden de la lógica que se necesita para describir el problema), han llevado a que los resultados obtenidos en alguno de estos campos sea transferido de manera inmediata al otro.

En uno de nuestros trabajos de investigación se ha introducido la definición de una restricción de $SO$, que consiste en limitar las relaciones que pueden tomar los cuantificadores de $SO$, considerando a la lógica como uno de los lenguajes de consulta a base de datos. El tipo de relaciones a los que estos cuantificadores pueden referirse son relaciones cerradas bajo $FO - type$. Esta lógica($SOF$) intenta lograr una lógica de mayor poder expresivo que la definida por Dawar($SO^w$) en la que los cuantificadores sólo pueden tomar relaciones cerradas bajo $FO - k$ tipos. Se demostró que nuestra lógica incluye estrictamente la de Dawar [4]. Se ha podido definir una nueva clase de complejidad descriptiva ($NPF$), que caracteriza el fragmento existencial de nuestra lógica gracias a modificar de las máquinas relacionales.

En otro de nuestros trabajos se estudia el impacto del aumento del orden de las variables en las lógicas. Se continúa con el estudio del poder expresivo de las lógicas $HO$ (High-Order logic) y en particular de los fragmentos de la lógica $VO$ (Variable-Order logic) definida en [11], que nace debido a que ninguna de las lógicas de orden superior cubre la clase completa de consultas computables ($CQ$)[1], es decir que no son completas. Además, si consideramos la unión de todas las lógicas de orden superior, es decir $\bigcup_{i\geq 2} HO^i$ ($HO^i$ representa la lógica de orden $i$), tampoco obtenemos una lógica completa. De aquí, se define $VO$ permitiendo el uso de variables de orden variable, mediante el uso de cuantificadores de orden. Las restricciones más importantes estudiadas sobre $VO$ son sobre: la cantidad de alternaciones de cuantificadores, la aridad de las variables de orden variable, los valores que pueden asignarse a las variables de orden en función del tamaño del dominio, el rango de cuantificadores y la cantidad de variables, de valuación y de orden.

## Resultados y Objetivos

Como trabajo futuro de esta línea de investigación se consideran varios aspectos relacionados al diseño de estructuras de datos que, consciente de la jerar-

quía de memorias y de las características particulares de los datos a ser indexados, saquen el mejor partido haciendólas eficientes en espacio y en tiempo.

Respecto de los lenguajes de consulta se continuará analizando la expresividad de distintas extensiones de FO y posibles restricciones de SO, para lograr caracterizar la clase de las consultas computables sobre bases de datos no convencionales.

En el caso de bases de datos métricas, se intentará que los índices se adapten mejor al espacio métrico particular considerado, por la determinación de su dimensión intrínseca, y también al nivel de la jerarquía de memorias donde se almacenará. Estos estudios sobre espacios métricos y sobre algunas estructuras de datos particulares permitirán no sólo mejorar el desempeño de las mismas sino también aplicar, eventualmente, muchos de los resultados que se obtengan a otros MAMs.

## Actividades de Formación

Dentro de esta línea de investigación se están formando alumnos y docentes-investigadores de acuerdo al siguiente detalle:

**Doctorado en Cs. de la Computación:** dos integrantes de la línea se encuentran desarrollando su tesis sobre la expresividad de la lógica como lenguaje de consulta. Otro integrante desarrolla su tesis sobre bases de datos métricas.

**Maestría en Cs. de la Computación:** un investigador de la línea está desarrollando su tesis en bases de datos métricas sobre búsqueda por similitud aproximada.

Además, se están dirigiendo actualmente tres trabajos finales de alumnos de la Licenciatura en Cs. de la Computación, prontos a finalizar.

## Referencias

[1] D. Harel A. K. Chandra. Computable queries for relational data bases. *J. of Computer and System Sciences*, 21(2):156–178, 1980.

[2] M. Barroso, N. Reyes, and R. Paredes. Enlarging nodes to improve dynamic spatial approximation trees. In *Procs. of the 3rd Int. Conf. on Similarity Search and Applications*, 41–48. ACM Press, 2010.

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

[4] A. Dawar. A restricted second order logic for finite structures. *Information and Computation*, 143:154–174, 1998.

[5] V. Dohnal, C. Gennaro, P. Savino, and P. Zezula. Similarity join in metric spaces. In *Proc. 25th European Conf. on IR Research*, LNCS 2634, 452–467, 2003.

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

[7] R. Fagin. Generalized first-order spectra and polynomial-time recognizable sets. *Complexity of Computation*, 7:43–73, 1974.

[8] J. Flum. H. Ebbinghaus. Finite model theory, second edition. *Springer*, 1999.

[9] N. Immerman. Descriptive and computational complexity. *Computational Complexity Theory*, 38:75–91, 1989.

[10] N. Immerman. Descriptive complexity. *Springer*, 1998.

[11] J. M. Turull Torres L. Hella. Computing queries with higher-order logics. *Theoretical Computer Science*, 355:197–214, 2006.

[12] G. Navarro. Searching in metric spaces by spatial approximation. *The Very Large Databases Journal (VLDBJ)*, 11(1):28–46, 2002.

[13] G. Navarro and N. Reyes. Dynamic spatial approximation trees. *J. of Experimental Algorithmics*, 12:1–68, 2008.

[14] G. Navarro and N. Reyes. Dynamic spatial approximation trees for massive data. In Tomás Skopal and Pavel Zezula, editors, *SISAP*, 81–88. IEEE Computer Society, 2009.

[15] V. Dohnal P. Zezula, G. Amato and M. Batko. Similarity search: The metric space approach. *Advances in Database Systems*, 32.

[16] R. Paredes. *Graphs for Metric Space Searching*. PhD thesis, University of Chile, 2008.

[17] 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*, LNCS 4007, 85–97, 2006.

[18] R. Paredes and N. Reyes. Solving similarity joins and range queries in metric spaces with the list of twin clusters. *J. of Discrete Algorithms*, 7(1):18–35, 2009.

[19] H. Samet. Foundations of multidimensional and metric data structures. *Morgan Kaufmann*, 2006.

[20] L. Stockmeyer. The polynomial-time hierarchy. *Theoret. Comput. Sci.*, 3:1–22, 1976.