Estructuras de Datos Métricas para la Recuperación de Información Multimedia en la Web. *

Roberto Uribe Paredes, Eduardo Peña Jaramillo
Departamento de Ingeniería en Computación
Universidad de Magallanes
Punta Arenas, Chile.

Grupo de Bases de Datos UART
Universidad Nacional de la Patagonia Austral
Río Turbio, Argentina.
(roberto.uribe@umag.cl, eduardo.pena@umag.cl)

and

Osiris Sofia
Universidad Nacional de la Patagonia Austral
Río Gallegos, Argentina.
(osofia@unpa.edu.ar)

Resumen

La *búsqueda por similaridad* consiste en recuperar todos aquellos objetos dentro de una base de datos que sean parecidos o relevantes a una determinada consulta. Este concepto tiene una amplia gama de aplicaciones en áreas como bases de datos multimediales, reconocimiento de patrones, minería de datos, recuperación de información, etc.

En este contexto, tres grupos de investigación aúnan esfuerzos en una misma dirección con el objetivo de posibilitar un mayor avance en torno al diseño, desarrollo e implementación de nuevas y eficientes estructuras métricas, así como también en la construcción de aplicaciones que permitan acercar a la realidad este tipo de investigaciones.

El presente artículo describe algunos de los avances realizados en el último año en torno a esta línea de investigación realizados por grupos conformados por la Universidad de Magallanes, Chile y por las unidades académicas de Río Turbio y Río Gallegos de Universidad Nacional de la Patagonia Austral, Argentina.

Palabras claves: bases de datos, estructuras de datos, algoritmos, espacios métricos, consultas por similitud, paralelismo, modelo BSP, CBIR.

*Este trabajo es parcialmente financiado por los proyectos y programas de investigación: PR-F1-02IC-08, UMAG, Chile; 29/C035 UNPA-UART y 29/A216 UNPA-UARG, Argentina.

1. Introducción

1.1. Antecedentes

Uno de los problemas de gran interés en ciencias de la computación es el de "búsqueda por similitud", es decir, encontrar los elementos de un conjunto más similares a una muestra. Esta búsqueda es necesaria en múltiples aplicaciones, como ser en reconocimiento de voz e imagen, compresión de video, genética, minería de datos, recuperación de información, etc. En casi todas las aplicaciones la evaluación de la similaridad entre dos elementos es cara, por lo que usualmente se trata como medida del costo de la búsqueda la cantidad de similaridades que se evalúan.

Interesa el caso donde la similaridad describe un espacio métrico, es decir, está modelada por una función de distancia que respeta la desigualdad triangular. En este caso, el problema más común y difícil es en aquellos espacios de "alta dimensión" donde el histograma de distancias es concentrado, es decir, todos los objetos están más o menos a la misma distancia unos de otros.

El aumento de tamaño de las bases de datos y la aparición de nuevos tipos de datos sobre los cuales no interesa realizar búsquedas exactas, crean la necesidad de plantear nuevas estructuras para

búsqueda por similaridad o búsqueda aproximada. Asimismo, se necesita que dichas estructuras sean dinámicas, es decir, que permitan agregar o eliminar elementos sin necesidad de crearlas nuevamente, así como también que sean óptimas en la administración de memoria secundaria. La necesidad de procesar grandes volúmenes de datos obligan a aumentar la capacidad de procesamiento y con ello la paralelización de los algoritmos y la distribución de las bases de datos.

Las distintas problemáticas mencionadas en el párrafo anterior son abarcadas por diferentes equipos, donde hay que incluir, además, el desarrollo y evaluación de prototipos de prueba.

1.2. Marco teórico

La similitud se modeliza en muchos casos interesantes a través de un espacio métrico, y la búsqueda de objetos más similares a través de una búsqueda por rango o de vecinos más cercanos.

Definición 1 (*Espacios Métricos*): Un espacio métrico es un conjunto X con una función de distancia $d : X^2 \rightarrow R$, tal que $\forall x, y, z \in X$,

1. $d(x, y) \geq 0$ *and* $d(x, y) = 0$ *ssi* $x = y$. *(positividad)*
2. $d(x, y) = d(y, x)$. *(Simetría)*
3. $d(x, y) + d(y, z) \geq (d(x, z)$. *(Desigualdad Triangular)*

Definición 2 (*Consulta por Rango*): Sea un espacio métrico *(X,d)*, un conjunto de datos finito $Y \subseteq X$, una consulta $x \in X$, y un rango $r \in R$. La consulta de rango alrededor de x con rango r es el conjunto de puntos $y \in Y$, tal que $d(x, y) \leq r$.

Definición 3 (*Los k Vecinos más Cercanos*): Sea un espacio métrico *(X,d)*, un conjunto de datos finito $Y \subseteq X$, una consulta $x \in X$ *y un entero k*. Los k vecinos más cercanos a x son un subconjunto A de objetos de Y, donde la $|A| = k$ y no existe un objeto $y \in A$ tal que *d(y,x)* sea menor a la distancia de algún objeto de A a x.

El objetivo de los algoritmos de búsqueda es minimizar la cantidad de evaluaciones de distancia realizadas para resolver la consulta. Los métodos para buscar en espacios métricos se basan principalmente en dividir el espacio empleando la distancia a uno o más objetos seleccionados. El no trabajar con las características particulares de cada aplicación tiene la ventaja de ser más general, pues los algoritmos funcionan con cualquier tipo de objeto [6].

Existen distintas estructuras para buscar en espacios métricos, las cuales pueden ocupar funciones discretas o continuas de distancia. Algunos son BKTree [4], MetricTree [14], GNAT [2], VpTree [19], FQTree [1], MTree [7], SAT [9], SlimTree [13], Spaghettis [5], SSS-Tree [3], EGNAT [15].

Algunas de las estructuras anteriores basan la búsqueda en pivotes y otras en clustering. En el primer caso se seleccionan pivotes del conjunto de datos y se precalculan las distancias entre los elementos y los pivotes. Cuando se realiza una consulta, se calcula la distancia de la consulta a los pivotes y se usa la desigualdad triangular para descartar candidatos.

Los algoritmos basados en clustering dividen el espacio en áreas, donde cada área tiene un *centro*. Se almacena alguna información sobre el área que permita descartar toda el área mediante sólo comparar la consulta con su centro. Los algoritmos de clustering son los mejores para espacios de alta dimensión, que es el problema más difícil en la práctica.

Existen dos criterios para delimitar las áreas en las estructuras basadas en clustering, *hiperplanos* y *radio cobertor* (*covering radius*). El primero divide el espacio en particiones de *Voronoi* y determina el hiperplano al cual pertenece la consulta según a qué centro corresponde. El criterio de radio cobertor divide el espacio en esferas que pueden intersectarse y una consulta puede pertenecer a más de una esfera.

1.3. Modelo de computación paralela BSP

El modelo BSP de computación paralela fue propuesto en 1990 con el objetivo de permitir que el desarrollo de software sea portable y tenga desempeño eficiente y escalable [18, 11]. BSP propone alcanzar este objetivo mediante la estructuración de la computación en una secuencia de pasos llamados *supersteps* y el empleo de técnicas aleatorias para el ruteo de mensajes entre procesadores. El computador paralelo, independiente de su arquitectura, es visto como un conjunto de pares procesadores-memoria, los cuales son conectados mediante una red de comunicación cuya topología es transparente al programador. Los supersteps son delimitados mediante la sincronización de procesadores. Los procesadores proceden al siguiente superstep una vez que todos ellos han alcanzado el final del superstep, los cuales son agrupados en bloques para optimizar la eficiencia de la comunicación. Durante un superstep, los procesadores trabajan asincrónicamente con datos almacenados en sus memorias locales. Cualquier mensaje enviado por un procesador está disponible para procesamiento en el

procesador destino sólo al comienzo del siguiente superstep. Dada la estructura particular del modelo de computación, el costo de los programas BSP puede ser obtenido utilizando técnicas similares a las empleadas en el análisis de algoritmos secuenciales. En BSP, el costo de cada superstep esta dado por la suma del costo en computación (el máximo entre los procesadores), el costo de sincronización entre procesadores, y el costo de comunicación entre procesadores (el máximo enviado/recibido entre procesadores).

2. Resultados Preliminares

Durante la primera etapa del trabajo conjunto entre los distintos grupos, básicamente se abordaron tres líneas de trabajo, la primera corresponde al rediseño de estructuras, su implementación y prueba. La segunda línea tiene relación con la paralelización de algoritmos y esquemas de distribución de estructuras sobre un cluster de PC's. Finalmente, la tercera corresponde a la implementación de aplicaciones.

El rediseño de estructuras ha sido orientada a aumentar la eficiencia en las bsquedas, es el caso de los trabajos [16, 17] como del diseño de nuevas estructuras [8]. En la actualidad se está trabajando sobre la optimización en memoria secundaria para el *SSS-Tree* y una nueva versión para la estructura *Lista de Cluster*, para ambos casos se espera presentar los resultados en congresos latinoamericos durante el presente año.

La paralelización se ha abordado sobre la estructura métrica *Spaghettis*, sobre ésta se han experimentado esquemas de distribución de datos sobre un cluster como la parelelización de sus algoritmos. También sobre esta estructura se han hecho modificaciones para darle características dinámicas, específicamente eliminación y reinserción, considerando en estos procesos el balance de la estructura en el cluster.

La tercera línea de trabajo está abocada a la implementación de aplicaciones, en este sentido, se implementó, en una etapa temprana, un *Digesto Digital Paralelo para Búsqueda por Similitud sobre Documentos* ([12]) y una segunda aplicación, ahora en etapa de prueba de inicial, se presenta en la siguiente subsección.

2.1. Sistema Recuperador de Imágenes Basado en Contenidos sobre Estructuras Métricas

Recuperar información desde una imagen basada en contenido (*CBIR: Content Based Image Retrieval*) corresponde a una metodología de recuperación con respecto al dominio de aplicación del proceso de recuperación en sí. Usa un análisis y procesamiento digital para generar descriptores a partir de los datos. Los méritos principales de sistemas basados en el contenido son: soporta el procesamiento de consultas visuales, la consulta es intuitiva y amistosa al usuario, la generación de los descriptores es automática, siendo objetiva y consistente.

El prototipo de prueba es una continuación del trabajo desarrollado en [10]. El prototipo está soportado por la estructura *EGNAT*, sin embargo, se realizaron experimentos del mismo *CBIR* sobre la estrutura *GNAT* y *Spaghettis*. En la figura 1 se puede observar resultados preliminares para un conjunto de consultas (primera columna) y los 5 primeros objetos recuperados. Este experimento fue realizado sobre una base de datos de 1,000 imágenes y un conjunto de 5 consultas.

3. Conclusiones

En este artículo se ha presentado una descripción breve del contexto y de algunos de los avances logrados por los equipos conformados por investigadores de las Universidades de Magallanes, Chile y Nacional de la Patagonia Austral, Argentina.

Los avances corresponden a trabajos realizados en conjunto en torno a la búsqueda por similitud en espacios métricos. Durante la primera mitad del presente año se espera enviar a evaluar estos resultados a congresos latinoamericanos.

Se espera continuar el trabajo con énfasis en el aumento de la eficiencia en las estructuras diseñadas y la continuación de pruebas sobre los prototipos indicados. Se espera contar, al finalizar el proyecto de investigación conjunto, con el desarrollo de parte de una máquina de búsqueda por similitud, soportada sobre un cluster de PCs, que pueda ser utilizada como prototipo en aplicaciones de tipo real.

Referencias

[1] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixedqueries trees. In *5th Combinatorial Pattern Matching (CPM'94)*, LNCS 807, pages 198–212, 1994.

[2] Sergei Brin. Near neighbor search in large metric spaces. In *the 21st VLDB Conference*, pages 574–584. Morgan Kaufmann Publishers, 1995.

[3] Nieves R. Brisaboa, Oscar Pedreira, Diego Seco, Roberto Solar, and Roberto Uribe. Clustering-based similarity search in metric spaces with sparse spatial centers. In

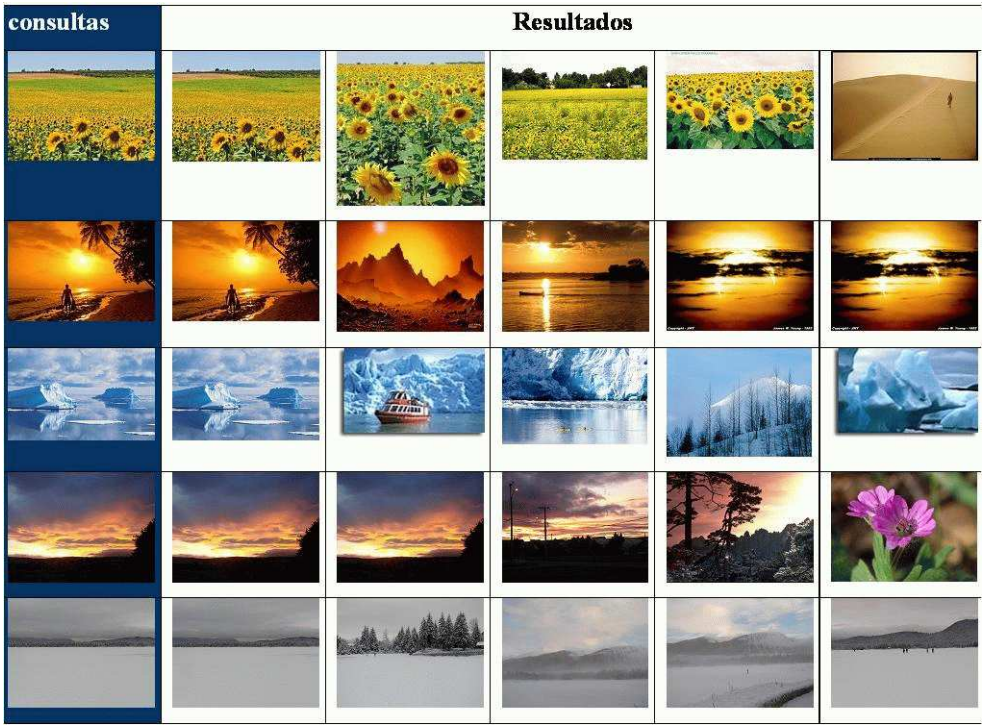

Figura 1: Primeras imágenes recuperadas por el *CBIR* para un conjunto de consultas (primera columna).

SOFSEM 2008: 34rd Conference on Current Trends in Theory and Practice of Computer Science, volume 4910 of *Lecture Notes in Computer Science*, pages 186–197, Novy Smokovec, High Tatras, Slovakia, January, 19-25 2008. Springer.

[4] W. Burkhard and R. Keller. Some approaches to best-match file searching. *Communication of ACM*, 16(4):230–236, 1973.

[5] E. Chavéz, J. Marroquín, and R. Baeza-Yates. Spaghettis: An array based algorithm for similarity queries in metric spaces. In *6th International Symposium on String Processing and Information Retrieval (SPIRE'99)*, pages 38–46. IEEE CS Press, 1999.

[6] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José L. Marroquín. Searching in metric spaces. In *ACM Computing Surveys*, pages 33(3):273–321, September 2001.

[7] P. Ciaccia, M. Patella, and P. Zezula. M-tree : An efficient access method for similarity search in metric spaces. In *the 23st International Conference on VLDB*, pages 426–435, 1997.

[8] Mauricio Marín, Veronica Gil-Costa, and Roberto Uribe. Hybrid index for metric space databases. In *Proc. of International Conference on Computational Science 2008 (ICCS 2008)*, volume 5101 of *Lecture Notes in Computer Science*, pages 327–336, Krakow, Poland, Jun 2008. Springer.

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

[10] Eduardo Peña-Jaramillo. Estructuras métricas paralelas en la recuperación de imágenes. Master's thesis, Escuela de Ingeniería, Departamento de Ciencias de la Computación, Pontificia Católica de Chile, Santiago, Chile, Nov. 2006.

[11] D.B. Skillicorn, J.M.D. Hill, and W.F. McColl. Questions and answers about BSP. Technical Report PRG-TR-15-96, Computing Laboratory, Oxford University, 1996. Also in *Journal of Scientific Programming*, V.6 N.3, 1997.

[12] Roberto Solar, Roberto Uribe-Paredes, Esteban Gesto, and Osiris Sofia. Implementación de un digesto digital paralelo para búsquedas por similitud sobre documentos. In *Congreso Argentino de Ciencias de la Computación*, La Rioja, Argentina, Octubre 2008. CACIC 2008.

[13] Caetano Traina, Agma Traina, Bernhard Seeger, and Christos Faloutsos. Slim-trees: High performance metric trees minimizing

overlap between nodes. In *VII International Conference on Extending Database Technology*, pages 51–61, 2000.

[14] J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. In *Information Processing Letters*, pages 40:175–179, 1991.

[15] Roberto Uribe-Paredes. Manipulación de estructuras métricas en memoria secundaria. Master's thesis, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile, Abril 2005.

[16] Roberto Uribe-Paredes, Claudio Márquez, and Roberto Solar. Estrategias de construcción sobre estructuras métricas para búsquedas por similitud. In *Conferencia Latinoamericana de Estudios en Informática (CLEI2008)*, Santa Fé, Argentina, 2008. CLEI 2008.

[17] Roberto Uribe-Paredes, Claudio Márquez, and Roberto Solar. Ssstree v2.0: Búsqueda por similitud en espacios métricos con solapamiento de planos. In *Congreso Argentino de Ciencias de la Computación*, La Rioja, Argentina, Octubre 2008. CACIC 2008.

[18] L.G. Valiant. A bridging model for parallel computation. *Comm. ACM*, 33:103–111, Aug. 1990.

[19] P. Yianilos. Data structures and algoritms for nearest neighbor search in general metric spaces. In *4th ACM-SIAM Symposium on Discrete Algorithms (SODA'93)*, pages 311–321, 1993.