Indexación y Recuperación de Información Multimedia

Jacqueline Fernández, Veronica Gil-Costa, Verónica Ludueña, Nora Reyes, Patricia Roggero
LIDIC, Departamento de Informática
Universidad Nacional de San Luis
{jmfer, gvcosta, vlud, nreyes, proggero}@unsl.edu.ar

Edgar Chávez
Escuela de Ciencias Físico–Matemáticas
Universidad Michoacana de San Nicolás de Hidalgo
elchavez@umich.mx

Resumen

En general, es difícil tanto para los usuarios que intentan recuperar información multimedia poder especificar claramente sus intereses y a través de una consulta bien definida, como para aquéllos que diseñan el sistema decidir cuáles de las características de los objetos multimedia pueden resultar relevantes. La forma en que los datos multimedia se representan, como ellos se almacenan y el costo de transferirlos, entre distintos niveles de la jerarquía de memoria o sobre una red, afectan directamente las respuestas del sistema. Dada una consulta al sistema, el objetivo clave de un sistema de recuperación de información es recuperar la información que podría ser útil o relevante para el usuario, en general haciendo uso de un índice especialmente diseñado para responder a las consultas de manera eficiente.

Así, nuestra línea de investigación está enfocada en lograr herramientas eficientes para recuperación de información multimedia, desarrollando nuevas técnicas capaces de soportar la interacción con el usuario, diseñando nuevas estructuras de datos (índices) capaces de manipular eficientemente datos multimedia y buscando representaciones para los datos que reflejen más adecuadamente las características de interés de los objetos multimedia.

Palabras Claves: *Recuperación de Información, Bases de Datos Multimedia, Indexación.*

Contexto

Esta línea de investigación se encuentra enmarcada dentro del Proyecto Consolidado 30310 de la Universidad Nacional de San Luis y en el Programa de Incentivos (código 22/F034): "Nuevas Tecnologías para el Tratamiento Integral de Datos Multimedia". Este proyecto es desarrollado en el ámbito del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) de la UNSL y se desarrollaron además dentro del Proyecto "Cooperación interuniversitaria para fortalecer el grado y postgrado en Procesamiento y Recuperación de Señales Digitales", dentro del Programa de Promoción de la Universidad Argentina para el Fortalecimiento de Redes Interuniversitarias III. Nuestra universidad participó junto con la Universidad de Zaragoza (España) y la Universidad Michoacana de San Nicolás de Hidalgo (México) (finalizado en 2010).

En este proyecto de investigación se pretende avanzar en la integración de las investigaciones sobre adquisición, preprocesamiento y análisis de datos no estructurados y su aplicación en dominios no convencionales. Se espera que el principal aporte de esta propuesta sea la incorporación de información no estructurada en los procesos de toma de decisiones y resolución de problemas que queda fuera de consideración en los enfoques clásicos.

Dentro de este contexto nuestra línea se dedica principalmente al diseño de índices que sirvan de apoyo a sistemas de recuperación de información orientados a datos no estructurados. Se espera así contribuir a estos sistemas obteniendo índices más eficientes para memorias jerárquicas, gracias a la compacticidad y/o con I/O eficiente. Se propone analizar las estructuras de datos existentes, y proponer nuevas u optimizaciones a las mismas, para manipular y recuperar algunos de los tipos de datos no estructurados que aparecen en entornos multimedia.

1. Introducción y Motivación

Es común en nuestros días que los sistemas de computación hagan uso intensivo de información estructurada, es decir datos elementales o estructuras, generadas con un formato específico por un programa determinado. Una característica principal en es-

tos casos, es que la estructura o formato de esta información puede ser fácilmente interpretada y directamente utilizada por un programa de computadora.

Si bien la información estructurada ha sido la principal materia prima utilizada en los sistemas computacionales hasta la fecha el hecho de restringirse al uso de este tipo de información conduce, muchas veces, a representar una visión parcial del problema y dejar fuera de consideración información que podría ser de gran importancia para la resolución efectiva del mismo. En este contexto gran parte de la información que se requiere para la toma de decisiones y la resolución de problemas de índole general proviene de información no estructurada, principalmente aquélla almacenada en forma de texto, audio, imagen y video.

Para responder eficientemente consultas sobre una base de datos multimedia se usan los métodos de acceso (índices) [12, 3, 10]. En general, en recuperación de información, se usan índices debido al volumen de datos con el que se trabaja.

Una forma de implementar la búsqueda por similitud en bases de datos multimedia es usando información de anotaciones que describan el contenido de los objetos multimedia. Sin embargo, este modo no es práctico en grandes repositorios multimedia porque las descripciones textuales deben generarse a mano, ya que son difíciles de obtener automáticamente. Además, en general, son subjetivas y en la mayoría de los casos no pueden caracterizar toda la información disponible del objeto multimedia. Más aún, otra restricción de este enfoque es que el procesamiento del índice debe "adivinar" cuáles clases de consultas se podrán hacer sobre los datos.

Un enfoque más prometedor para implementar sistemas de recuperación usando búsqueda por similitud es una búsqueda basada en contenidos, la cual usa el dato multimedia mismo. Para calcular la similitud entre dos objetos multimedia, se debe definir una función de distancia. Dicha función mide la similitud, o más bien la disimilitud, entre dos objetos.

El concepto de búsqueda por similitud se puede definir a partir del concepto de espacios métricos, lo cual da un marco formal que es independiente del dominio de aplicación. Un espacio métrico está compuesto por un *universo* $\mathcal{U}$ de objetos y una función de distancia $d : \mathcal{U} \times \mathcal{U} \longrightarrow R^+$, que satisface las propiedades que hacen de ella una métrica. Las consultas por similitud, sobre una base de datos $\mathcal{S} \subseteq \mathcal{U}$, son usualmente de dos tipos:

Búsqueda por rango: recuperar todos los elementos de $\mathcal{S}$ a distancia r de un elemento q dado.

Búsqueda de los k vecinos más cercanos: dado q, recuperar los k elementos más cercanos a q en $\mathcal{S}$.

Si la base de datos $\mathcal{S}$ posee n objetos, las consultas pueden ser trivialmente respondidas llevando a cabo n evaluaciones de distancia. Sin embargo, la mayoría de las aplicaciones requieren distancias costosas de computar, por ejemplo, la comparación de huellas digitales, búsquedas por contenido en bases de datos multimedia, etc. Por lo tanto, la búsqueda secuencial no escala para problemas de tamaño medio o grande, que son los tamaños más habituales de las bases de datos multimedia. Así el objetivo es preprocesar la base de datos, construyendo un índice, para que las consultas puedan ser respondidas con la menor cantidad de cálculos de distancia.

Un caso particular de dato multimedia son las imágenes. Una imagen es un arreglo de píxeles que resultan de la convolución de una función señal y de una función de "rendering". Dos imágenes son consideradas copia una de otra si tienen la misma señal, aunque tengan diferente función de "rendering". El estudio del conjunto de características para una recuperación exitosa de imágenes basadas en su contenido representa un aspecto fundamental en muchas aplicaciones tales como reconocimiento de objetos, recuperación de imágenes, reconocimiento de texturas entre otras. El color, la forma o la textura son características visuales de la imagen importantes para analizar, representar e indexar la imagen [5, 6, 7]. Una imagen puede sufrir distorsiones geométricas como RST (rotación, escalado y traslación), "cropping", perspectiva; distorsiones de iluminación ("light distortion", "sunburnt distortion"), así como también distorsiones de calidad ("blurring" o "halftone scan-line"). En sistemas típicos los contenidos visuales de las imágenes son extraídos y descriptos mediante vectores característicos multidimensionales. Los vectores deben ser unívocos y robustos a las transformaciones a las que puede estar sometida una imagen. Existe una extensa literatura proponiendo diferentes vectores o descriptores y comparando la performance entre ellos. Dependiendo de la aplicación, algunos descriptores son mas apropiados que otros.

Por lo tanto, nuestra propuesta se enfoca en tratar de mejorar las herramientas de recuperación desarrollando nuevas técnicas que soporten la interacción con el usuario, diseñando nuevas estructuras de

datos (índices) capaces de manipular eficientemente datos multimedia y, dado que es fundamental la etapa de extracción automática de características de interés, definiendo representaciones que reflejen más adecuadamente los objetos multimedia.

2. Líneas de Investigación

Como se ha mencionado previamente se pretende investigar respecto de dos aspectos importantes para los sistemas de recuperación de información multimedia: diseñar nuevos índices capaces de manipular eficientemente datos multimedia y definir representaciones que reflejen las características de interés de los objetos multimedia, gracias a la obtención de un descriptor que sea robusto frente a operaciones (por ejemplo el cizallado).

Búsqueda en Texto Comprimido Auto-indexado

Los tiempos de respuesta de las consultas dentro de una fracción de segundo en los motores de búsqueda Web son factibles debido a la utilización de las técnicas de indexación y el almacenamiento en caché, que están pensados para grandes colecciones de textos particionado y replicado en un conjunto de procesadores de memoria distribuida. En este contexto se estudian métodos de procesamiento de consultas alternativas para esta configuración, que se basa en una combinación de texto comprimido auto-indexado y listas de posteo (posting lists). Un texto auto-indexado (es decir, un índice que comprime el texto y es capaz de extraer las piezas arbitrarias del mismo) puede ser competitivo con un índice invertido si tenemos en cuenta el proceso de consulta completo, que incluye la descompresión del índice, el ranking y el tiempo de extracción de *snnipets*. La ventaja es que en el espacio de la colección de documentos comprimidos, se puede llevar a cabo la generación de listas de posteo, el ranking de documentos y extracción de snnipets. Esto reduce significativamente el número total de procesadores que participan en la solución de las consultas.

Diseño de Índices

Un catálogo importante de índices para espacios métricos aparece en [10, 3, 12]. La mayoría de los índices usan la desigualdad triangular para evitar el análisis secuencial de la base de datos. La distancia entre la consulta y los objetos de la base de datos puede ser estimada calculando algunas distancias de antemano hacia unos objetos distinguidos llamados *pivotes* y sin calcular las distancias reales desde el objeto de consulta a los objetos de la base de datos durante una búsqueda. Otra técnica común es indexar una partición del espacio en regiones denominadas *particiones compactas*.

Entre todas las técnicas para indexación en espacios métricos nos interesan las estructuras de datos dinámicas, donde la base de datos no se conoce de antemano y además tanto los objetos como las consultas arriban al azar. En cambio, las estructuras estáticas se benefician desde el conocimiento de la base de datos seleccionando los mejores puntos de referencia para una estructura de datos determinada.

En particular es de interés mejorar el desempeño de índices dinámicos jerárquicos (árboles), que es el caso de algunos de los índices para espacios métricos. Estos índices dinámicos, en general, se construyen incrementalmente vía inserciones. De tal manera, la raíz del árbol es el primer objeto que llega, y esto se repite recursivamente en cada nivel del árbol.

En esta línea se ha propuesto una técnica donde el "buffering" logra un buen compromiso entre una estructura estática construída con toda la información necesaria y una dinámica con conocimiento local de los datos. Entonces, en lugar de elegir al primer elemento como la raíz, se demora la selección hasta que hayan arribado suficientes elementos para estar en condiciones de realizar dicha selección, y de esta manera se toma una decisión en base a más información. Dado que las consultas arriban a un ritmo desconocido, para mantener el dinamismo es necesario contar con un índice que responda a las consultas con mejor desempeño que la técnica de fuerza bruta. La idea ha sido, entonces, dar una estructura propia al "buffer" de manera que fuera capaz de responder consultas. Es por ello que el índice del "buffer" debería ser rápido y eficiente. En este sentido, se han analizado dos elecciones posibles: la primera fue usar un índice del mismo tipo como estructura del "buffer", reconstruyendo una vez que el "buffer" estuviera completo. La segunda alternativa fue usar otra estructura de datos, como *AESA* [11], donde se asume el conocimiento completo de las distancias entre los elementos del "buffer". Así, se han analizado distintas estrategias de selección de la raíz.

Esta técnica provee un marco adecuado para diseñar estructuras de datos dinámicas estables. Por lo tanto, tener un "buffer" en todos los niveles de una estructura jerárquica debería ser útil cuando se di-

señan estrategias de ruteo para guiar las búsquedas, lo cual resulta un área promisoria de investigación. Resultados preliminares fueron publicados en [4].

En muchos casos los volúmenes de información con los que se debe trabajar (millones de imágenes en la Web), hacen necesario que los índices sean almacenados en memoria secundaria. En este caso, para hacerlos eficientes, no sólo se debe considerar que durante las búsquedas se realice el menor número de cálculos de distancia sino también, dado el costo de las operaciones sobre disco, se efectúe la menor cantidad posible de operaciones de E/S. Por ello, en esta línea nos hemos dedicado a diseñar índices especialmente adaptados para trabajar en memoria secundaria, logrando un buen desempeño de los mismos, principalmente en las búsquedas.

Hemos diseñado e implementado las siguientes estructuras *DSACL*-tree* y el *DSACL+-tree* [2], las cuáles son optimizaciones para memoria secundaria de la estructura propuesta en [1] y demostraron ser competitivas frente a otras de las estructuras para memoria secundaria conocidas tales como el *M-tree* y *DSA*-tree* y *DSA+-tree* [8]. Nos proponemos optimizarlas aún más, gracias a la aplicación de técnicas de computación de alto desempeño.

Representación de Objetos Multimedia

La performance es cada vez más importante en visión por computadora. Para implementar sistemas de recuperación efectivos usando búsqueda por similitud se realiza una búsqueda basada en contenidos, la cual usa el dato multimedia mismo y se define una función de distancia para calcular la similitud entre dos objetos multimedia.

En muchos casos para modelar la similitud de objetos multimedia se transforman los objetos en puntos de un espacio vectorial, el cual es un tipo particular de espacio métrico. Luego de definir ciertas características de interés para los objetos, se extraen los valores numéricos que los representa y se construye el vector de características o descriptor, generalmente de alta dimensionalidad. Sobre espacios vectoriales se han definido numerosas funciones de distancia; por ejemplo: la distancia Euclidiana. El tipo de aplicación, las características a explotar o la dimensionalidad son aspectos fundamentales a considerar para definir la mejor función de distancia a utilizar. Por lo tanto, es necesario resolver un problema de optimización.

En el caso de los espacios métricos, la función de similitud (la distancia) normalmente mide el mínimo esfuerzo (costo) necesario para transformar un objeto en otro. Aunque ésta es una manera formal de definir la similitud entre objetos, dependiendo de los tipos de datos multimedia reales la función de similitud puede ser muy compleja y puede no siempre satisfacer las propiedades de una métrica.

Dependerá realmente de la aplicación final cuál de ambos modelos debería usarse. El enfoque más práctico, como ya se mencionó, es modelar los datos multimedia como un espacio vectorial. Sin embargo, el enfoque de espacios métricos puede ser también útil si la similitud satisface las propiedades de una métrica. Cuando los datos multimedia se han modelado como un espacio métrico o como un espacio vectorial, la búsqueda por similitud se reduce a una búsqueda de objetos o puntos cercanos en el espacio.

3. Resultados

Se ha podido comprobar experimentalmente que las estrategias de "buffering" mejoran la performance de una estructura de datos dinámica [4]. Se seleccionó como caso de estudio el Árbol de Aproximación Espacial Dinámico (DSA-tree) [8] y se obtuvo una mejora sistemática en los costos de las consultas usando un "buffer" en el primer nivel del árbol.

En particular, se ha podido verificar que esta estructura es mejor que su versión estática [8], porque deja como "vecinos" de un nodo algunos objetos alejados, permitiendo así avanzar en la exploración espacial con "pasos más grandes" a los vecinos de un nodo. Es por esto que ahora se pretende analizar el efecto de elegir como vecinos de un nodo una muestra de objetos cercanos y lejanos. Si se clasificaran los objetos por distancia a la raíz, usando la información del histograma de distancias a ella, se podría elegir con esa misma densidad a los vecinos: muchos donde el histograma sea denso y pocos donde el histograma sea ralo. Se espera que esta estrategia mejore, aún más, el desempeño de este índice y que esto pueda extenderse a otros índices jerárquicos.

En este mismo sentido, se implementaron dos versiones de índices (DSACL*-tree y DSACL+-tree) que permiten trabajar con grandes volúmenes de datos, por haber sido diseñadas específicamente para trabajar en memoria secundaria y que mostraron ser competitivas contra otras estructuras igualmente diseñadas para tal fin [2]. Se espera lograr para estos índices una implementación paralela eficiente.

Respecto a la representación de datos multimedia, en particular sobre imágenes, un problema abierto

corresponde a la construcción de una distancia que permita identificar imágenes completas [9]. Es decir, el estado del arte ([5, 6, 7]) permite identificar puntos de interés que son semejantes; pero es sólo un paso para poder identificar imágenes completas que contienen muchos puntos de interés. Por lo tanto, para una imagen se desea llegar a determinar los puntos de interés de la misma, que sean invariantes a distorsiones. Entonces, estos puntos de interés para una imagen formarían su representación, donde los diferentes puntos de interés serían a su vez vectores. Luego, la recuperación basada en contenido considerará, como imágenes similares, a aquellas imágenes cuyos grafos que se forman con los respectivos conjuntos de puntos de interés sean similares. Así uno de los objetivos consistirá en determinar una función de distancia entre los respectivos grafos de puntos de interés.

4. Formación de Recursos Humanos

Considerando la importancia de la formación, para contribuir al desarrollo de sistemas de recuperación de información multimedia, dentro de esta línea se están formando los siguientes docentes:

Trabajo Final de Licenciatura en Ciencias de la Computación: un alumno desarrolló su trabajo final sobre un índice dinámico para búsquedas por similitud en espacios métricos especialmente diseñado para memoria secundaria, gracias a una Beca Estímulo de la Facultad de Ciencias Físico Matemáticas y Naturales de la UNSL. El mismo continuará sus estudios de Maestría profundizando en la misma temática.

Tesis de Maestría en Ciencias de la Computación: uno de los integrantes está desarrollando su tesis de Maestría sobre el tema de recuperación de imágenes.

Tesis de Doctorado en Ciencias de la Computación: uno de los integrantes se encuentra definiendo su plan de doctorado sobre temas de diseño y optimización de índices para realizar búsquedas por similitud, con miras a aplicaciones de minería de datos multimedia.

Referencias

[1] M. Barroso, N. Reyes, and R. Paredes. Enlarging nodes to improve dynamic spatial approximation trees. In *Proc. of the 3rd International Conference on Similarity Search and Applications (SISAP 2010)*, pages 41–48. ACM Press, 2010.

[2] L. Britos, M. Printista, and N. Reyes. Dynamic spatial approximation trees with clusters for secondary memory. *Anales del XVI Congreso Argentino de Ciencias de la Computación (CACIC 2010)*, 712–721, 2010.

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

[4] E. Chávez, N. Reyes, and P. Roggero. Delayed insertion strategies in dynamic metric indexes. In M. Arenas and B. Bustos, editors, *SCCC*, 34–42. IEEE Computer Society, 2009.

[5] H. Ling and D. Jacobs. Deformation invariant image matching. In *ICCV '05: Proc. of the Tenth IEEE International Conference on Computer Vision*, 1466–1473, 2005. IEEE.

[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[7] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

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

[9] A. Ocsa, A. Rodriguez, H. Chuctaya, and G. Humpire. Reconocimiento de rostros mediante puntos característicos locales. In *Anales del II Simposio Peruano de Computación Gráfica y Procesamiento de Imgágenes*, 1–6, 2008.

[10] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann Publishers Inc., 2005.

[11] E. Vidal. An algorithm for finding nearest neighbors in (approximately) constant average time. *Pattern Recognition Letters*, 4:145–157, 1986.

[12] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag, 2005.