Procesamiento de Consultas y Aplicaciones del Modelo Métrico Temporal

Anabella De Battista , Andrés Pascal, Alejandra Díaz, Pablo Gancharov, Adrian Planas

Departamento de Sistemas de Información, Fac. Reg. Concepción del Uruguay

Universidad Tecnológica Nacional, Entre Ríos, Argentina

{debattistaa, pascalj}@frcu.utn.edu.ar

{alejandraalciradiaz,pablogancharov,pladnic}@gmail.com

Norma Edith Herrera

Departamento de Informática

Univ. Nac. de San Luis , San Luis, Argentina

nherrera@unsl.edu.ar

Gilberto Gutierrez

Facultad de Ciencias Empresariales

Universidad del Bio-Bio, Chillán, Chile

ggutierr@ubiobio.cl

Resumen

Las bases de datos actuales permiten almacenar datos no estructurados tales como imágenes, sonido, video, datos geométricos, etc. Las tecnologías tradicionales de bases de datos no son aplicables en este ámbito. Nuevos modelos de bases de datos surgen para cubrir las necesidades de almacenamiento y búsqueda en estas bases de datos. Entre estos nuevos modelos podemos nombrar el espacial, el temporal, el de espacios métricos y el métrico-temporal, entre otros. Nuestro área de investigación es el diseño de índices eficientes para estos nuevos modelos de bases de datos.

Palabras Claves: Espacios Métricos, Bases de Datos Métrico-Temporales, Índices

1. Contexto

El presente trabajo se desarrolla en el ámbito del Grupo de Investigación en Bases de Datos (PID 25-D040) perteneciente al Departamento de Sistemas de la Universidad Tecnológica Nacional, F. R. Concepción del Uruguay, cuyo objetivo principal es el estudio de métodos de acceso, procesamiento de consultas y aplicaciones de bases de datos no tradicionales.

2. Introducción

Las bases de datos clásicas se organizan bajo el concepto de búsqueda exacta sobre datos estructurados. Esto significa que la información se organiza en registros los cuales se dividen en campos que contienen valores completamente comparables. Una consulta retorna todos aquellos registros cuyos campos coinciden con los aportados en la consulta (búsqueda exacta). Una característica importante de las bases de datos clásicas es que capturan sólo un estado de la realidad modelizada, usualmente el más reciente. Por medio de las transacciones, la base de datos evoluciona de un estado al siguiente descartando el estado previo.

Actualmente las bases de datos han inclui-

do la capacidad de almacenar datos no estructurados tales como imágenes, sonido, texto, video, datos geométricos, etc. La problemática de almacenamiento y búsqueda en estos tipos de base de datos difiere de las bases de datos clásicas en varios aspectos: los datos no son estructurados por lo que no es posible organizarlos en registros y campos; la búsqueda exacta carece de interés; resulta de interés mantener todos los estados de la base de datos y no sólo el más reciente para poder consultar el intervalo de tiempo de vigencia de los objetos. Es en este contexto donde surgen nuevos modelos de bases de datos.

El modelo de ***espacios métricos*** [5], permite trabajar con objetos no estructurados y realizar búsquedas por similitud sobre los mismos. Un espacio métrico es un par (U, d) donde U es un universo de objetos y $d : U \times U \rightarrow R^+$ es una función de distancia definida entre los elementos de U que mide la similitud entre ellos. Una de las consultas típicas en este modelo es la búsqueda por rango, denotado por $(q, r)_d$, que consiste en recuperar los objetos de la base de datos que se encuentren como máximo a distancia r de un elemento q dado.

El modelo de ***bases de datos temporales*** [16] incorpora al tiempo como una dimensión, por lo que permite asociar tiempos a los datos almacenados y consultar por los objetos vigentes en un intervalo o en un instante de tiempo dado.

Existen aplicaciones donde resulta de interés realizar búsquedas por similitud teniendo en cuenta también la componente temporal. Es en este ámbito donde surge el ***modelo métrico temporal***. En este modelo se puede trabajar con objetos no estructurados con tiempos de vigencia asociados y realizar consultas por similitud y por tiempo en forma simultánea. Formalmente un *Espacio Métrico-Temporal* es un par (U, d), donde $U = O \times N \times N$, y la función d es de la forma $d : O \times O \rightarrow R^+$. Cada elemento $u \in U$ es una triupla (obj, t_i, t_f), donde obj es un objeto (por ejemplo, una imagen, sonido, cadena, etc) y $[t_i, t_f]$ es el intervalo de vigencia de obj. La función de distancia d, que mide la similitud entre dos objetos, cumple con las propiedades de una métrica (positividad, simetría y desigualdad triangular). Una *consulta métrico-temporal* se define como una 4-upla $(q, r, t_{iq}, t_{fq})_d$, tal que $(q, r, t_{iq}, t_{fq})_d = \{o/(o, t_{io}, t_{fo}) \in X \wedge d(q, o) \leq r \wedge (t_{io} \leq t_{fq}) \wedge (t_{iq} \leq t_{fo})\}$.

.

3. Líneas de Investigación

Nuestra principal línea de trabajo es el estudio de métodos de acceso, procesamiento de consultas y aplicaciones de bases de datos no tradicionales, centrándonos principalmente en el modelo métrico-temporal. Damos a continuación una descripción de las líneas de investigación que actualmente estamos desarrollando.

3.1. Índices en Memoria Secundaria

Varios índices métrico-temporales se han propuesto en este ámbito: el *FHQT-Temporal* [11], el *Historical-FHQT* [3], el *Event-FHQT* [10] y el *Pivot-FHQT* [2]; todos ellos han tomado como base el Fixed Height Queries Tree[1], un índice para espacios métricos.

Los índices desarrollados hasta el momento se basan en el supuesto de que la memoria principal tiene capacidad suficiente como para mantener tanto el índice como la base de datos. Si esto no es así, la cantidad de accesos a memoria secundaria realizados durante el proceso de búsqueda es un factor crítico en la performance del índice.

En [6] se presenta el *Compact Pat Tree*, un índice en memoria secundaria para búsquedas de patrones en texto. Básicamente este índice consiste en una representación compacta de un árbol binario (un Pat-Tree) en disco. En [14]

se presenta una modificación de esta técnica de paginación para árboles r-arios, la cual es fácilmente adaptable para la paginación de índices métricos-temporales.. En el caso particular del FHQT-Temporal la aplicación de esta técnica requiere de pocas modificaciones. Nos encontramos desarrollando la implementación del FHQT-Temporal en disco.

3.2. Consultas Métrico Temporales sobre Cadenas

Los índice desarrollados hasta el momento han sido evaluados empíricamente con lotes generados a partir de imágenes del sitio SISAP (http://sisap.org/Home.html), añadiendo a cada imagen un intervalo de vigencia. Nuestro interés es probar la eficiencia de los índices en un ámbito real de uso.

Para ello hemos desarrollado una aplicación que tiene por finalidad permitir efectuar consultas métrico-temporales sobre el sistema de archivos de Windows/Linux, dependiendo en donde se ejecute. Esta aplicación está orientada a la búsqueda de archivos y carpetas tanto por nombre como por fecha, con diferentes radios de búsqueda, mediante la generación de los distintos tipos de índices métrico-temporales.

El objetivo es probar la eficiencia de los índices desarrollados sobre una base de datos de cadenas (nombres de archivos/carpetas) con una cantidad importante de elementos (unos $250{,}000$ nombres) y que sea de utilidad real. El planteo de nuestra solución esta dada para una aplicación totalmente amigable para el usuario que desconoce el tema, donde sólo debe elegirse a partir de qué directorio se realizará la búsqueda. Nos encontramos en la etapa de evaluación experimental de la aplicación, a fin de probar la eficiencia de los distintos índices métrico-temporales en este ámbito de uso.

3.3. Búsqueda de Imágenes

En la búsqueda de imágenes por similitud en grandes bases de datos, es tan importante la eficiencia del sistema (recuperar imágenes en un tiempo razonable) como su eficacia (recuperar imágenes que sean realmente de interés). La eficacia depende principalmente del preprocesamiento de las imágenes, de la técnica de extracción de características y de la función de distancia que se emplee. Por otro lado, los factores de mayor relevancia para la eficiencia del proceso son el costo de la función de distancia y el tipo de índice que se utilice para acelerar la búsqueda.

La mayoría de los sistemas comerciales existentes para recuperación de imágenes asocian texto a cada una de las imágenes y transforman el problema de búsqueda de imágenes en el problema de búsqueda en texto. Pero en muchas aplicaciones esto no es suficiente. Un verdadero sistema de recuperación de imágenes debe permitir dar una imagen como objeto de consulta y debe poder determinar la similitud entre ese objeto y cada una de las imágenes de la base de datos en forma eficiente, a fin de responder la consulta. Estos sistemas se denominan Content-based Image Retrieval Systems (CBIR) [7, 9, 18], y están en pleno desarrollo.

La búsqueda por similitud aplicada a imágenes implica transformar las imágenes en vectores de características de esas imágenes, para luego insertar esos vectores en un índice métrico. Luego, ante una consulta, se debe transformar la imagen de consulta de la misma manera en que se transformaron las imágenes de la base de datos para poder proceder a buscar usando el índice.

Existen dos tareas que son cruciales en este proceso: una es convertir las imágenes en vectores; la otra es definir una función de distancia que permita comparar las imágenes. La primera tarea afecta directamente la eficacia del sistema, dado que las búsquedas se realizarán en base a las características extraídas

de cada imagen. La segunda tarea afecta tanto la eficacia como la eficiencia; la eficacia porque la función de distancia modela formalmente lo que se entiende por similitud y la eficiencia porque el costo de búsqueda en el índice se ve directamente afectado por el costo de cálculo de la función de distancia y por la distribución de distancias que genera.

Si bien hay numerosos trabajos de investigación que se concentran en el preprocesamiento de las imágenes y extracción de características [8, 17], las funciones de distancia [15, 13] y los índices métricos [4, 5], la mayoría lo hace por separado, sin estudiar la integración de estos aspectos.

En este linea hemos trabajado definiendo un proceso para el tratamiento integral de una base de datos de imágenes. El proceso propuesto consta de una etapa de preprocesamiento de la base de datos y otra posterior para la realización de consulta. La etapa de preprocesamiento tiene dos objetivos: uno escalcular la paleta de colores reducida y adaptada a las imágenes de la base de datos; el otro es calcular los histogramas de colores de las imágenes, en base a la paleta definida en el punto anterior.

Una vez obtenida la paleta de colores, se procesan las imágenes calculando los histogramas mediante la transformación de cada color en el más cercano de la paleta. Para clasificar cada pixel se utiliza la distancia euclideana entre colores. Posteriormente se construye un índice métrico utilizando los histogramas como vectores característicos de cada imagen, y empleando alguna función de distancia. En nuestro trabajo hemos usado las siguientes funciones de distancia: Euclideana, Manhattan, Euclideana por Cuadrantes y Manhattan por Cuadrantes (en ambos casos con 16 cuadrantes). Las evaluación experimental de las funciones de distancia han mostrado que la distancia euclideana por cuadrantes posee un buen comportamiento tanto en la calidad de sus resultados como en la dimensionalidad intrínseca de espacio métrico generado [12]. Actualmente estamos trabajando en la inclusión de funciones de distancia que consideren otros aspectos tales como textura o forma.

4. Resultados Esperados

Se espera contar con métodos eficientes, tanto en memoria principal como en memoria secundaria, para el procesamiento de consultas en el ámbito de bases de datos no tradicionales. Esto incluye el diseño de índices, la definición de funciones de distancias adecuadas a la problemática tratada, la definición de nuevas consultas que sean de interés y el desarrollo de aplicaciones en ámbitos reales de uso de los métodos desarrollados.

5. Formación de Recursos Humanos

El trabajo desarrollado hasta el momento forma parte del desarrollo de dos Tesis de Maestría en Ciencias de la Computación, una de ellas fue defendida y aprobada en marzo del 2009, y la otra tiene fecha de finalización estimada para agosto del presente año. Uno de los integrantes del grupo está desarrollando su Tesis Doctoral sobre la temática de indexación en memoria secundaria de bases de datos textuales, que está íntimamente relacionado a la temática de estudio de este grupo. El grupo cuenta en la actualidad con tres alumnos becarios que se están formando en estas temáticas.

Referencias

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

[2] A. De Battista, A. Psacal, N. Herrera, and G. Gutierrez. Metric-temporal access methods. *Journal of Computer Science & Technology,*, 10(2):54–60, 2010.

[3] De Battista, A. Pascal, G. Gutierrez, and N. Herrera. Un nuevo índice métrico-temporal: el historical fhqt. In *Actas del XIII Congreso Argentino de Ciencias de la Computación*, Corrientes, Agentina, 2007.

[4] E. Chávez and K. Figueroa. Faster proximity searching in metric data. In *Proceedings of MICAI 2004. LNCS 2972, Springer*, Cd. de México, México, 2004.

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

[6] D. Clark and I. Munro. Efficient suffix tree on secondary storage. In *Proc. 7th ACM-SIAM Symposium on Discrete Algorithms*, pages 383–391, 1996.

[7] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008.

[8] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval - a quantitative comparison. In *In DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, pages 228–236, 2004.

[9] L. R. Long, S. Antani, T. Deserno, and G. Thoma. Content-based image retrieval in medicine. *International Journal of Healthcare Information Systems and Informatics*, 4(1):1–16, 2009.

[10] A. Pascal, A. De Battista, G. Gutierrez, and N. Herrera. Indice metrico-temporal event-fhqt. In *Actas del XIIII Congreso Argentino de Ciencias de la Computación*, La Rioja, Argentina, 2008.

[11] A. Pascal, De Battista, G. Gutierrez, and N. Herrera. Procesamiento de consultas métrico-temporales. In *XXIII Conferencia Latinoamericana de Informática*, pages 133–144, Costa Rica, 2007.

[12] A Planas, A. Pascal, A. De Battista, Alejandra Díaz, and Norma Herrera. Métodos de acceso para bases de datos métrico - temporales. In *Actas del I Seminario Argentina - Brasil de Tecnologías de la Información y la Comunicación*, Rosario, Argentina, 2011.

[13] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the International Conference on Computer Vision-Volume 2*, Washington, DC, USA, 2001. IEEE Computer Society.

[14] D. Ruano, N. Herrera, C. Ruano, and A. Villegas. Representación en memoria secundaria del trie de sufijos. In *Actas del XVI Congreso Argentino de Ciencias de la Computación*, Buenos Aires, 2010.

[15] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

[16] B. Salzberg and V. J. Tsotras. A comparison of access methods for temporal data. *ACM Computing Surveys*, 31(2), 1999.

[17] A. Shahbahrami and D. B. Juurlink. Comparison between color and texture features for image retrieval. In *Proceedings of the 19th Annual Workshop on Circuits, Systems and Signal Processing*, 2008.

[18] R. C. Veltkamp and M.Tanase. Content-based image retrieval systems: A survey. Technical report, UU-CS-2000-34, 2000.