Filtro efectivo para la búsqueda por similaridad utilizando permutaciones en un espacio métrico

Karina Mariela Figueroa Mora[1] y Roxana Reyes[2]

1 Fac. de cs físico-matemáticas, 2 Fac. de cs. Físico-matemáticas. karina@fismat.umich.mx

Hoy en día existe una gran cantidad de datos digitales multimedia en los que es indispensable realizar búsquedas para obtener información. El problema se complica cuando lo que andamos buscando son *resultados parecidos.* Por ejemplo, piense en un retrato hablado de un criminal, si se quisiera conocer quiénes son las personas dentro de una base de datos que se parecen al del retrato el tipo de búsqueda que se debe realizar es la búsqueda por similaridad. Precisamente, este tipo de búsqueda consiste en recuperar los elementos mas parecidos dentro de una base de datos a una consulta. En lo sucesivo de manera general nos referiremos a objetos en la base de datos. La similaridad deberá ser representada por una función de distancia que mida que tanto se parece una imagen a otra. Dado que tenemos una base de datos y una función de distancia el problema puede ser modelado como un espacio métrico. La metodología es: primero se creará un índice que permita las búsquedas de manera rápida. Después se medirá la eficiencia del índice propuesto en la fase de consultas.

Existen diferentes algoritmos para la búsqueda por similaridad en este tipo de espacios. El mas reciente y competitivo es el conocido como *algoritmo basado en permutaciones.* El cual consiste en: seleccionar un conjunto objetos de la base de datos, éstos serán llamado permutantes. Los objetos restantes deben medir su distancia hacia los permutantes y ordenarlas de manera ascendente, dicho orden será conocido como permutación. El conjunto de permutaciones de todos los objetos será llamado *índice.*

Al momento de tener una consulta, primero, ésta debe crear su permutación y posteriormente buscar las permutaciones similares en la base de datos, note que esto es proceso secuencial pues debe revisar todas las permutaciones. Una vez revisadas todas, tendremos una lista de objetos candidatos a ser parte de la respuesta.

Nuestra propuesta consiste en emplear un filtro que evite el proceso secuencial y reduzca el conjunto de candidatos a ser parte de la respuesta. El filtro presentado identifica las posibles permutaciones con mayor probabilidad de formar parte de la repuesta. Nuestros resultados muestran que es posible reducir a solo una fracción de objetos y no a un proceso secuencial.

La fase experimental se ha llevado acabo con diversas bases de datos como son: imágenes obtenidas de Flickr, imágenes obtenidas de la NASA, y un diccionario de palabras en inglés y español.

En el póster presentaremos algunos conceptos básicos, la metodología del trabajo y por supuesto por medio de un ejemplo, mostrar el algoritmo que emplearemos como filtro para mejorar el tiempo de respuesta en la búsqueda por similaridad.