USO DEL MÉTODO DE COMPRESIÓN Re-Pair PARA DISMINUIR EL ESPACIO REQUERIDO POR LOS ÍNDICES UTILIZADOS EN ESPACIOS MÉTRICOS BASADOS EN PERMUTACIONES.

FELIPE IGNACIO PINO DURAN
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

En bases de datos de objetos provenientes de un espacio métrico, las consultas de elementos se hacen por medio del método de búsqueda por proximidad o similitud entre los elementos. Para esto es necesaria la indexación de estas bases de datos. El índice basado en permutaciones (IBP) resulta muy eficaz, con respecto a otros algoritmos de búsqueda por proximidad, para resolver este tipo de consultas. Pero en la práctica, los IBPs requieren utilizar una excesiva cantidad de espacio en memoria principal. Por consiguiente, el objetivo principal de esta memoria es comprimir estos índices aplicando el método de compresión Re-Pair.

La compresión con Re-Pair consiste en reemplazar los pares de símbolos que más se repiten en las permutaciones con nuevos símbolos simples. Para lograr esto, se implementa un algoritmo recursivo con estructuras de datos tales como Heaps Binarios y Tablas de Hash, con el fin de lograr una eficiente compresión, dado que por lo general se tienen que revisar millones de símbolos.

El método propuesto se compara con la alternativa de compactar los IBPs. Esto consiste en guardar la representación binaria de cada IBP en una variable de máquina. La concatenación de la secuencia de bits formada representa un número Considerado como ´índice compacto.

Las pruebas experimentales se realizan con bases de datos vectoriales de baja y alta dimensionaldad. En donde se mide y compara el rendimiento y el grado compresión de los algoritmos. Re-Pair resulta ser muy efectivo para dimensionalidades bajas (en efecto, comprime hasta en un 92% en dimensión). El peor escenario evaluado, que corresponde a dimensión 1024, el IBP se comprime a un 46% del espacio original, mientras que el algoritmo compacto lo reduce a un 25%, pero tiene la desventaja que la des compactación del IBP es mucho más lenta que su descompresión .Para terminar, es necesario señalar que reducir el espacio utilizado por el IBP en sí es bueno porque permite trabajar con volúmenes de datos más grandes sin tener que usar espacio en disco como memoria virtual.

Hay que recordar que esto último mermaría considerablemente en el rendimiento de las consultas por el costo que emplearía el sistema operativo al realizar la paginación de memoria principal y virtual.

Palabras claves: Espacio Métrico, ´Índice basado en Permutaciones, Re-Pair.

ABSTRACT

In object database from a metric space, queries made through elements of the search method by proximity or similarity between elements. This requires indexing such databases. The permutation-based index (IBP) is very e↵ective, with respect to other proximity search algorithms to solve such queries. But in practice, use IBPs require an excessive amount of space in main memory. Therefore, the main objective of this document is to compress these indexes using the compression methodRe-Pair. Compression with Re-Pair is to replace pairs of symbols that are repeated in permutations with new simple symbols. To accomplish this, a recursive algorithm is implemented using data structures such as Hash Tables and Heaps Binary, in order to achieve efficient compression, since they usually have to review million symbols. The proposed method is compared with the alternative of compact IBPs. This is to save the binary representation of each IBP variable machine. Concatenating the bit string formed represents an number considered compact index. Experimental testing with vector data bases low and high dimensionality. Where it is measured and compared to performance and degree of compression algorithms. Re-Pair is very e↵ective for low dimensionality (in fact compressed to 92% in dimension). The worst scenario evaluated, corresponding to dimension 1024, IBP is compressed to 46 % of the original space while the compact algorithm reduces to 25 %, but has the disadvantage that the de-compaction IBP is much slower its decompression. Finally, it should be noted that reducing the space used by the IBP itself is good because it allows you to work with bigger data volumes without using disk space as virtual memory. Remember that the latter significantly undermine the performance of queries for the cost that it would use the operating system to paging the main memory and virtual memory.

Keywords: Metric Space, Permutation based Index, Re-Pair.

.