CONSTRUCCIÓN DE GRAFOS DE K VECINOS EN ESPACIOS METRICOS DE ALTA DIMENSIONALIDAD

ALEJANDRO ANTONIO SAZO ROJAS
INGENIERO CIVIL EN COMPUTACION

RESUMEN

Sea U un conjunto de elementos y d una función de distancia entre ellos. Sean NNk(u) los k elementos con la distancia más pequeña en U n fug de acuerdo a la función d. El grafo de k vecinos más cercanos (kNNG por sus siglas en inglés) es un grafo dirigido con peso G(U;E) que conecta a cada uno de los elementos con sus k vecinos más cercanos, es decir, E = f(u; v); 8u 2 U; v 2 NNk(u)g. La construcción del kNNG es una generalización directa del problema de todos los vecinos más cercanos (ANN por sus siglas en inglés), así que ANN corresponde al problema de construir el 1NNG. Los kNNGs son una parte central en muchas aplicaciones: detección de clusters y datos aislados, diseño VLSI (Very Large Scale Integration) y otras simulaciones de procesos físicos, reconocimiento de patrones, sistemas de consulta o recomendación de documentos, entre otros. Hay muchos algoritmos de construcción de kNNGs que asumen que los nodos son puntos en RD y d es la distancia Euclidiana o alguna distancia de la familia de distancias de Minkowski. Sin embargo, ese no siempre es el caso en muchas aplicaciones de los kNNGs. En esta memoria se proponen dos algoritmos que usan un índice llamado Lista de Clusters (tanto la implementación tradicional, a menudo llamada LC como también la versión dinámica, a menudo llamada RLC), el cual es un índice métrico basado en particiones compactas. Se presentan resultados analíticos y experimentales para evaluar el comportamiento de los algoritmos y se comparan los resultados obtenidos con el algoritmo de construcción básico. El éxito de la Lista de Clusters (LC=RLC) en dimensiones altas radica en el hecho que cambia tiempo de construcción por tiempo de consulta. Esto se ve reflejado en los resultados experimentales. En efecto, la evaluación experimental muestra que los algoritmos tienen costos de la forma c1n1:12 para espacios métricos de baja y media dimensionalidad y c1n1:41 para espacios métricos de alta dimensionalidad.

Palabras claves: kNNGs, espacios métricos.

ABSTRACT

Let U be a set of elements and d a distance function among them. Let NNk(u) be the k elements in U n fug having the smallest distance to u according to the function d. The k-Nearest Neighbor Graph (kNNG) is a weighted directed graph G(U;E) connecting each element to its k-nearest neighbors, thus E = f(u; v); 8u 2 U; v 2 NNk(u)g. Building the kNNG is a direct generalization of the all-nearest- neighbor problem, so ANN corresponds to the 1NNG construction problem. KNNGs are central in many applications: cluster and outlier detection, VLSI design, spin glass and other physical process simulations, pattern recognition, query or document recommendation systems, and others. There are many kNNG construction algorithms which assume that nodes are points in RD and d is the Euclidean or some Lp Minkowski distance. However, this is not the case in several kNNG applications. In this work we propose two algorithms that use an index called List of Clusters (the traditional one, often called just LC for short and the dinamic version, often called just RLC for short), which is a metric space index based on compact partitions. We present analytical and experimental results to evaluate their behavior and compare our results with the basic construction algorithm. The key to success of the List of Clusters (LC=RLC) in high dimensions is that it trades construction time for query time. This can be appreciated in the experimental evaluation. As a matter of fact, our experimental results show that our algorithms have costs of the form c1n1:12 distance computations for low and medium dimensionality spaces, and c1n1:41 for high dimensionality spaces.

keywords: kNNGs, metric spaces.