Tarea 2

CC5213 – Recuperación de Información Multimedia

13 de Noviembre de 2018

El objetivo de esta tarea es estudiar la relación entre efectividad y eficiencia en las búsquedas aproximadas del vecino más cercano y estudiar características que influyen en su desempeño.

Se han calculado descriptores para el contenido audiovisual de los videos utilizados en la Tarea 1. Específicamente se tienen tres pares de conjuntos **Q** y **R**, donde para cada vector de **Q** se desea encontrar su vecino más cercano en **R**, es decir, dada una función de distancia **d** para todo vector **q** de **Q** se desea determinar el vector **r** de **R** tal que $\mathbf{d}(\mathbf{q},\mathbf{r}) \leq \mathbf{d}(\mathbf{q},\mathbf{x})$ para todo vector **x** de **R**.

Para evaluar las búsquedas aproximadas, primero debe localizar el vecino más cercano real usando el algoritmo **Linear Scan** o de fuerza bruta. Luego, con los vectores de **R** debe construir los siguientes índices[1]:

- **Randomized KD-Tree** con distintas cantidades de árboles (`trees`).
- **K-Means Tree** con distintas cantidades de centroides por nivel (`branching`).

Cada índice permite resolver búsquedas aproximadas del vecino más cercano con distintos valores de aproximación (`checks`). Para cada valor de aproximación la búsqueda aproximada logra cierta **Efectividad** (fracción de consultas cuya respuesta es correcta[2]) y **Eficiencia** (fracción del tiempo de búsqueda comparado con Linear Scan). Notar que con esta definición el algoritmo **Linear Scan** obtiene efectividad 1 y eficiencia 1 mientras que una buena búsqueda aproximada obtendría un resultado cercano a efectividad 1 y eficiencia 0.

Para cada uno de los tres conjuntos de descriptores realice las siguientes tareas:

1. Calcule el histograma de distancias del conjunto **R**, es decir, grafique la distribución de los valores de $\mathbf{d}(\mathbf{x},\mathbf{y})$ para una muestra aleatoria de **x** e **y** en **R**, y determine su dimensión intrínseca ($\rho=\mu^2/2\sigma^2$).
2. Determine el mejor índice para resolver búsquedas aproximadas del vecino más cercano. Para esto, construya distintos índices y realice búsquedas aproximadas con distintos valores de aproximación. Tabule los resultados obtenidos y elabore un gráfico con las **curvas de efectividad versus eficiencia**.

---

[1] Una implementación de estos índices (incluído Linear Scan) se encuentra en la librería FLANN (*Fast Library for Approximate Nearest Neighbors*). El módulo de Python es `pyflann`.

[2] Notar que puede existir más de una respuesta correcta para el vecino más cercano.

3. Analice cómo cambia el desempeño del mejor índice encontrado en el punto anterior al utilizar **Análisis de Componentes Principales (PCA)**. Para esto, use PCA para reducir **Q** y **R** a distintas dimensionalidades y evalúe el desempeño del índice en cada una. Tabule los resultados obtenidos y elabore un gráfico con las **curvas de efectividad versus eficiencia**. Notar que la efectividad se mide con respecto a los vecinos más cercanos en el espacio original (sin PCA).

Analice los resultados obtenidos para los tres conjuntos de descriptores y responda las siguientes preguntas:

a) ¿Qué tipo de índice permite obtener la mejor performance en búsquedas aproximadas? ¿Es posible usar algún criterio para determinar sus mejores parámetros (`trees`, `branching`, `checks`)? ¿Qué sucede si se considera el tiempo de construcción?

b) ¿Cómo es afectado el desempeño de las búsquedas aproximadas de un índice al usar PCA sobre los vectores de búsqueda?

c) ¿Es posible usar algún criterio para predecir la efectividad y eficiencia que tendrán las búsquedas aproximadas en un conjunto **R** dado?

Construya un reporte de Jupyter Notebook (.ipynb) que responda estas preguntas, incluyendo los procesos de carga de datos y generación de gráficos. Para mejorar la legibilidad del código puede incluir archivos .py externos.

Puede usar módulos comunes de Python para análisis de datos, como matplotlib, numpy, scipy, sklearn. **Debe usar Python 3**.

Se evaluará la correctitud de los experimentos realizados, la calidad de los gráficos generados, las conclusiones obtenidas y su sustento.

Los datos los puede descargar desde la dirección: http://juan.cl/CC5213-2018b/

El plazo máximo de entrega es el **Martes 27 de noviembre a las 23:59** por U-Cursos. No incluya datos de prueba. La tarea es ***individual***.