Algoritmos para Big Data

García Prado, Sergio
sergio@garciparedes.me

16 de marzo de 2017

Resumen

*En este documento se expone una breve descripción acerca de las distintas disciplinas de estudio relacionadas con el ámbito del tratamiento de grandes cantidas de información (*Big Data*) desde una perspectiva algorítmica.*

1. Introducción

El procesamiento de cantidades masivas de información presenta un gran reto a nivel computacional debido a su elevado coste originado por el gran tamaño en la entrada. Para solventar dicha problemática, se prefieren algoritmos que posean como característica un orden de complejidad sublineal ($o(N)$), tanto en tiempo como en espacio. Dichas técnicas se llevan a cabo sobre paradigmas de computación paralela, que permiten aprovechar en mayor medida las restriciciones físicas a nivel de hardware.

2. Algoritmos para Streaming

Los *Algoritmos para Streaming* se caracterizan por procesar los datos de una forma secuencial dependiente del orden de llegada. La ventaja que presentan respecto de otras alternativas en tiempo real es la utilización de propiedades estadísticas (se enmarcan por tanto, dentro de los *Algoritmos Probabilísticos*) para reducir su coste, lo que añade una determinada tasa de error. El descubrimiento de métodos altamente eficientes para estimar *Momentos de Frecuencia* ha sido un gran hito dentro de esta categoría algorítmica.

3. Estructuras de Datos de Resumen

Para reducir el coste derivado de la obtención de resultados valiosos sobre conjuntos masivos de datos, es necesario apoyarse en diferentes estructuras que resuman su información, de manera que el coste de procesamiento a partir de estas estructuras sea mucho más asequible. Se utilizan sobre conjuntos de datos de distínta índole, como *streamings en tiempo real*, *bases de datos estáticas* o *grafos*. Existen distintas técnicas como *Sampling*, *Histogram*, *Wavelets* o *Sketch*. A continuación se realiza una breve descripción acerca de esta última.

3.1. Sketch

Son estructuras de datos que se basan en la idea de procesar el conjunto completo de datos de entrada aplicando la misma operación sobre cada una de las instancias (lo que permite su uso en entornos tanto estáticos como dinámicos) para almacenar caracteríscas de de las mismas. Destacan los *Sketches lineales*, que permiten su generación de manera distribuida. Para mantener estas estructuras se utilizan *Algoritmos para Streaming*, puesto que se enmarcan perfectamente en dicho contexto. Los *Sketches* permiten la obtención de propiedades estadísticas referentes al conjunto de datos. Entre las distintas alternativas destacan *Count-Sketch*, *CountMin-Sketch*, *AMS Sketch*, *HyperLogLog*, etc.

4. Redución de la Dimensionalidad

Los algoritmos que utilizan técnicas de reducción de dimensionalidad se basan en la intuicción del lema de *Johnson–Lindenstrauss*, que demuestra la existencia de funciones para la redución de la dimensión del espacio con un porcentaje de distorsión mínima. Estas técnicas son utilizadas en algoritmos para la *busqueda de los vecinos más cercanos*, la *multiplicación aproximada de matrices* o el aprendizaje mediante *Manifold Leaning*.

5. Paralelización a gran Escala

El paradigma de alto nivel sobre el que se lleva a cabo el procesamiento de conjuntos de datos de gran escala se apoya fuertemente en técnicas de paralelización. La razón se debe al elevado tamaño de la entrada, que no permite su almacenamiento en la memoria de un único sistema.

5.1. Modelo MapReduce

El modelo *MapReduce* ha sufrido un crecimiento exponencial en los últimos años debido a su alto grado de abstracción, que oculta casi por completo cuestiones relacionadas con la implementación de bajo nivel al desarrollador y su capacidad para ajustarse a un gran número de problemas de manera eficiente.

6. Técnicas de Minería de Datos

Una de las razones por las cuales es necesaria la investigación de nuevos algoritmos de carácter sublineal es la necesidad de obtención de información valiosa a partir de conjuntos masivos de datos. A este fenómeno se le denomina *Minería de Datos*. Existen dos grandes categorías denominadas: *Clasificación* (determinar una clase de pertenencia) y *Regresión* (determinar un valor continuo). Para ello existen distintas técnicas como: *Árboles de Decisión*, *Métodos Bayesianos*, *Redes Neuronales*, *Máquinas de Vector Soporte*, *Manifold Leaning*, etc.