

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 áreas del conocimiento relacionadas con el tratamiento de grandes cantidades 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 un elevado coste originado por el gran tamaño en la entrada. Para solventar dicha problemática, se prefieren algoritmos que posean un orden de complejidad sublineal ($o(N)$) sobre todo en espacio. Dichas técnicas se llevan a cabo sobre paradigmas de computación paralela, lo que permite aprovechar en mayor medida las restricciones a nivel de hardware.

2. ALGORITMOS PARA STREAMING

Los *Algoritmos para Streaming* se caracterizan por procesar las instancias del conjunto de datos secuencialmente e imponen como restricción que el orden de dicha operación sea irrelevante para el resultado final. La ventaja que presentan respecto de otras alternativas en tiempo real, como los *Algoritmos Online*, es la utilización de propiedades estadísticas (se enmarcan por tanto, dentro de los *Algoritmos Probabilísticos*) para reducir su coste, lo que por contra, añade una determinada tasa de error. El descubrimiento de métodos altamente eficientes para estimar los *Momentos de Frecuencia* ha marcado 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 los sintetizen, de manera que el coste de procesamiento a partir de estas estructuras se convierta en una tarea mucho más asequible. Se utilizan sobre conjuntos de datos de distinta 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 técnica.

3.1. SKETCH

Son estructuras de datos que se basan en la idea de realizar sobre cada una de las instancias del conjunto de datos la misma operación (lo que permite su uso en entornos tanto estáticos como dinámicos) para recolectar distintas características. Destacan los *Sketches lineales*, que permiten su procesamiento de manera distribuida. Para mantener estas estructuras se utilizan *Algoritmos para Streaming*, puesto que se encajan perfectamente en el contexto descrito. Los *Sketches* permiten realizar distintas preguntas sobre propiedades estadísticas referentes al conjunto de datos. Los ejemplos más destacados son: *Count-Sketch*, *CountMin-Sketch*, *AMS Sketch*, *HyperLogLog*, etc.

4. REDUCCIÓN DE LA DIMENSIONALIDAD

Los algoritmos que utilizan técnicas de reducción de dimensionalidad se basan en la intuición originada a partir del lema de *Johnson-Lindenstrauss*, que demuestra la existencia de funciones para la reducción de la dimensión espacial con un ratio de distorsión acotado. Estas técnicas son utilizadas en algoritmos para la *busqueda de los vecinos más cercanos*, la *multiplicación aproximada de matrices* o el aprendizaje mediante *Manifold Learning*.

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, se utilizan distintas técnicas como: *Árboles de Decisión*, *Métodos Bayesianos*, *Redes Neuronales*, *Máquinas de Vector Soporte*, *Manifold Learning*, etc.