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 grandes cantidades de información presenta un gran reto a nivel computacional debido al elevado coste originado por el elevado tamaño del conjunto de datos. Para solventar dicha problemática se prefieren algoritmos cuya principal característica es 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 actuales 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 estos presentan respecto de otras alternativas en tiempo real es que utilizan propiedades estadísticas para reducir su coste, lo que añade una cierta tasa de error. El descubrimiento de métodos altamente eficiente para estimar *Momentos de Frecuencia* ha sido un gran hito en esta categoría algorítmica.

3. Estructuras de Datos de Resumen

Para reducir el coste necesario para obtener resultados valiosos del conjunto masivo de datos es necesario apoyarse en diferentes estructuras de datos que resumen la información presente en el mismo, de manera que la complejidad de procesamiento a partir de estas sea asumible.

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 tanto en entornos estáticos como en tiempo real) para almacenar caracteríscas de de las mismas. Para dicha labor se utilizan *Algoritmos para Streaming*, puesto que se enmarcan perfectamente en dicho contexto. Estas estructuras permiten la obtención de propiedades estadísticas del conjunto de datos. Entre las distintas alternativas destacan *Count-Sketch*, *CountMin-Sketch*, *AMS Sketch*, *HyperLogLog*...

4. Redución de la Dimensionalidad

[TODO]

5. Paralelización a gran Escala

[TODO]

5.1. Modelo MapReduce

[TODO]

6. Técnicas de Minería de Datos

Una de las razones por las cuales es necesaria la investigación sobre algoritmos como los citados anteriormente es la obtención de nuevos resultados a partir de conjuntos masivos de datos. A este fenómeno se le denomina *Minería de Datos*. Desde el punto de vista de la obtenición de un nuevo atributo referido a un determinado dato, hay dos grandes categorías que se corresponden con la *Clasificación* (determinar el grupo al que pertenece) y la *Regresión* (determinar el valor numérico que tomará). Para ello existen distintas alternativas como *Árboles de Decisión*, *Métodos Bayesianos*, *Redes Neuronales*, *Máquinas de Vector Soporte*...