Algoritmos para Big Data:

Grafos y PageRank

Sergio García Prado

E.T.S. Ingeniería Informática, UVa

Índice General

Introducción

Big Data

Los algoritmos para **Big Data** son aquellos que se encargan de resolver problemas sobre conjuntos de datos de tamaño masivo.

Problema de Accesos a Memoria

[TODO]

Soluciones a la complejidad del Big Data

- Algoritmos para Streaming
- Técnicas de Reducción de la Dimensionalidad
- Estrategias de Paralelización
- ...

Algoritmos para Streaming

Modelo en Streaming

- Serie Temporal:
  $1, 5, 3, -4, 2, -3, 5, ...$
- Caja Registradora (Cash-Register):
  $(2, +1), (3, +4), (1, +3), (2, +3), (4, +5), ...$
- Molinete (Turnstile):
  $(2, -1), (3, -4), (1, +3), (2, -3), (4, +5), ...$

Algoritmos para Streaming

Los **Algoritmos para Streaming** son aquellos que procesan la entrada de manera secuencial, teniendo en cuenta únicamente el elemento actual, junto con una estimación de los procesados anteriormente, utilizando un orden sublineal $o(n)$ en espacio respecto del rango de posibles valores en la entrada.

Algoritmos para Streaming

$$F_k = \sum_{i=1}^{n} m_i^k \tag{1}$$

- Algoritmo de Morris: F_1
- Algoritmo de Flajolet-Martin: F_0
- Estimación de Momentos de Frecuencia: $F_k,\ k \in N^*$

Algoritmo de Flajolet-Martin

[TODO]

Estrategias de Sumarización

Estrategias de Sumarización

- Muestreo Aleatorio
- Histogramas
- Wavelets
- Sketches

Sketches

- Count-Min Sketch
- Count Sketch
- AMS Sketch
- Hyper-LogLog
- L_p-Samplers

Count-Min Sketch

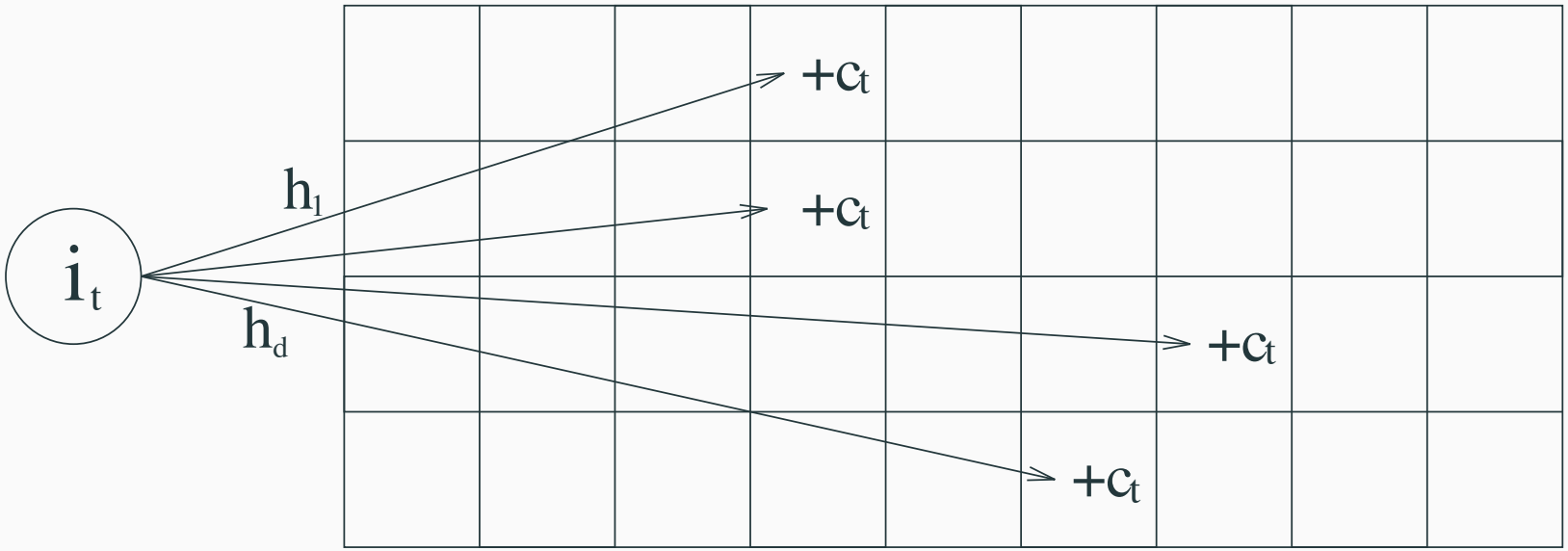


Figura 1:

Algoritmos aplicados a Grafos

Algoritmos aplicados a Grafos

Sea $G = (V, E)$ un grafo formado por $n = |V|$ vértices y $m = |E|$ aristas, de tal manera que $e_i = (v_{i_1}, v_{i_2}) \in E$ y $\{v_{i_1}, v_{i_2}\} \in V$

Sobre el **Modelo en Semi-Streaming** se procesa un grafo a través del stream de aristas, en un espacio poli-logarítmico respecto del cardinal de vértices utilizando un número reducido de pasadas sobre el stream.

Spanners y Sparsifiers

- Spanner:
- Sparsifier:

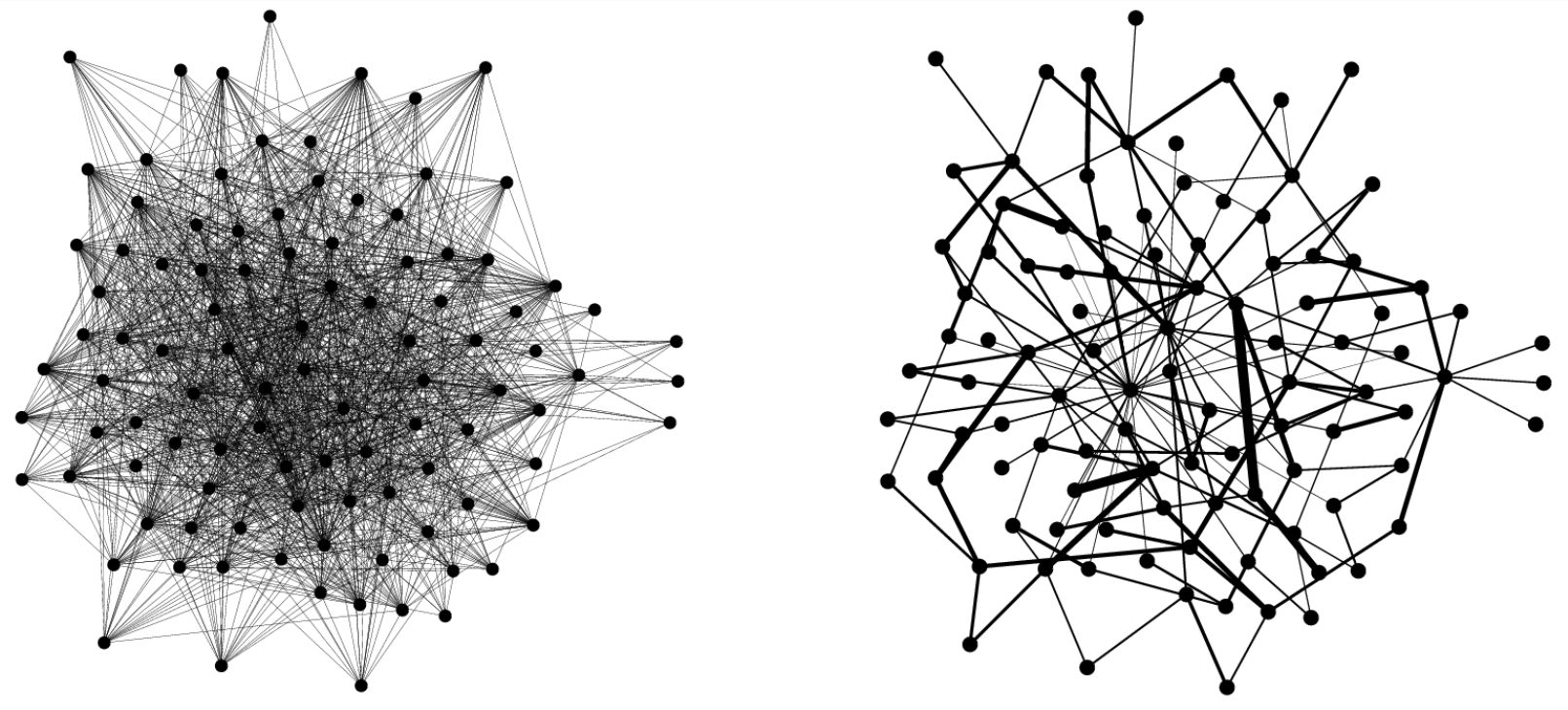


Figura 2:

Problemas sobre Grafos

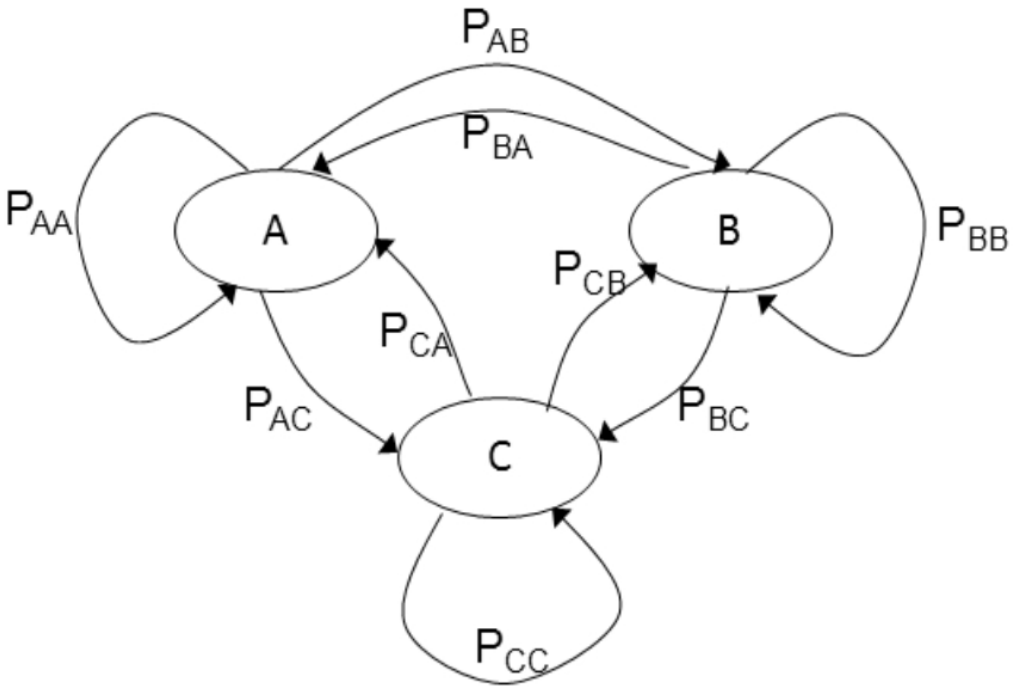
- Verificación de Grafo Bipartito
- Conteo de Triángulos
- Árbol Recubridor Mínimo
- Componentes Conectados
- Paseos Aleatorios
- ...

Algoritmo PageRank

Algoritmo PageRank

[TODO]

Cadena de Markov



$$\mathbf{P}=\begin{bmatrix} p_{AA} & p_{AB} & p_{AC} \\ p_{BA} & p_{BB} & p_{BC} \\ p_{CA} & p_{CB} & p_{CC} \end{bmatrix}$$

Figura 3:

¿Cómo se calcula?

- Algebraica [TODO]
- Iterativa [TODO]
- Paseos Aleatorios

PageRank Personalizado

[TODO]

Algoritmos Similares

- HITS
- SALSA
- SimRank

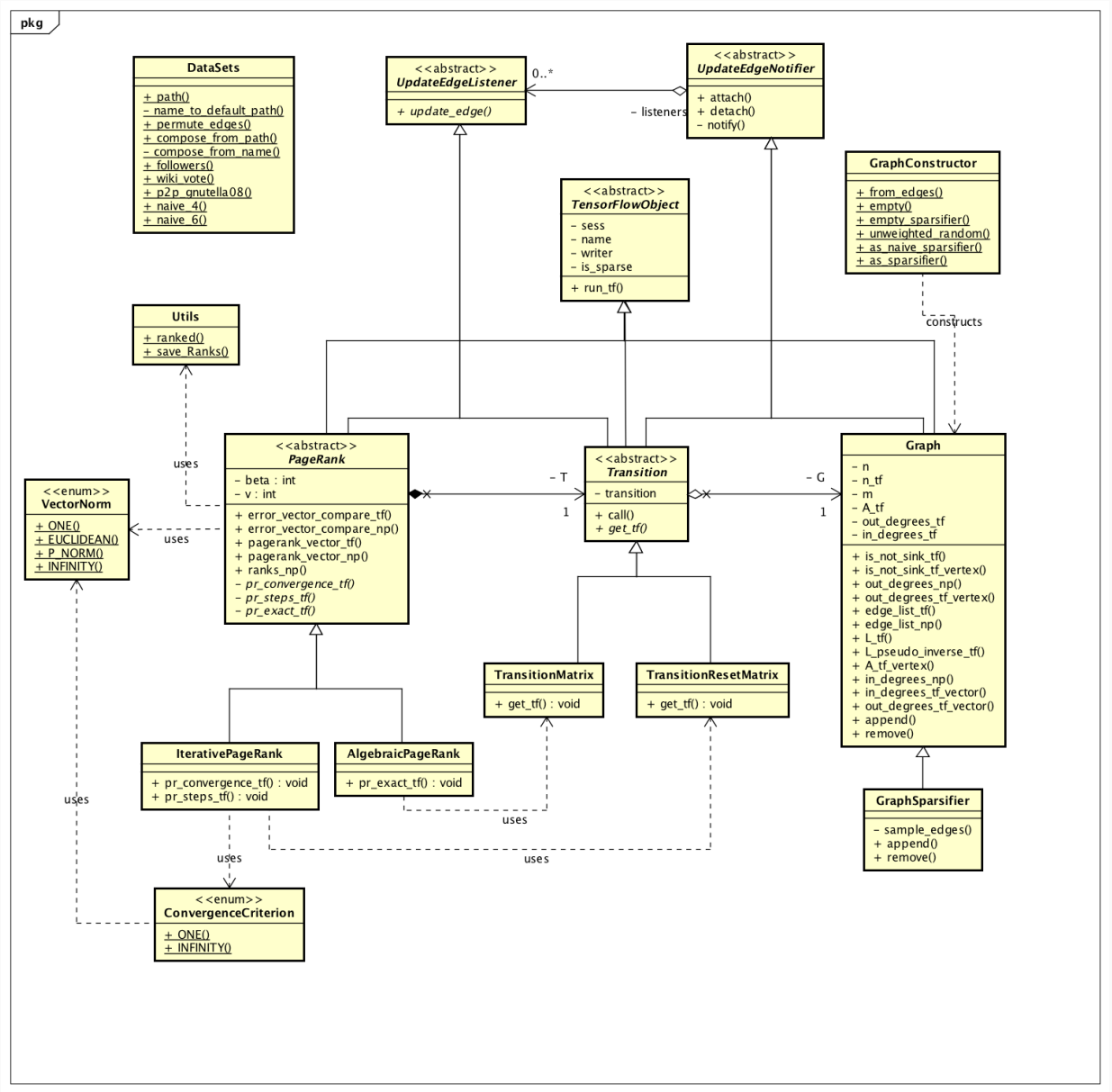
Implementación

Implementación

La implementación realizada consiste en una biblioteca de grafos (**tf_G**) utilizando como base la plataforma de cálculo matemático intesivo **TensorFlow**.

tf_G se encuentra en una fase muy temprana, formando únicamente el conjunto de métodos para el cálculo del **PageRank**.

Diagrama de Clases



¿Preguntas?

Referencias i