UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

Titulo: Indexado de Wikipedia a través de una arquitectura Map-Reduce
Autores: José Luis Larroque
Director: Alicia Diaz
Codirector: Diego Torres
Asesor profesional: -
Carrera: Licenciatura en Informática, plan 2003

Resumen

Se realizó un algoritmo que permite generar un indice de caminos entre dos artículos cualesquiera de Wikipedia. Este algoritmo fue desarrollado adaptando Wikipedia para ser procesada como un grafo en Giraph, un framework de procesamiento de grafos utilizado por grandes compañías como Facebook, Twitter, etc. La plataforma de computo utilizada para ejecutar este trabajo fue Amazon Web Services y Elastic Map Reduce, a través de una subvención para investigación. La arquitectura utilizada para el trabajo fue Map Reduce.

Palabras Claves

Giraph – Java – Map Reduce – Cloud Computing- Amazon Web Services

Conclusiones

Se logró el objetivo de poder generar los caminos posibles entre dos artículos cualesquiera de Wikipedia. Se debe continuar con el desarrollo para adaptar el algoritmo a contextos mas desafiantes, como grafos mas densos.

Trabajos Realizados

Se investigó como construir localmente un cluster en Hadoop, así como desarrollar algoritmos que corran en el. Se investigó el funcionamiento de Giraph,de forma de poder realizar algoritmos en este framework. Se desarrolló un algoritmo capaz de buscar caminos en un grafo del tamaño de Wikipedia (versión en español), usando como tecnologías principales el framework Giraph (el cual está preparado para correr en arquitecturas Map Reduce). Se probó el mismo en la plataforma de Cloud Computing Amazon Web Services. Se documentaron los resultados de estas evaluaciones y se los analizó.

Trabajos Futuros

Optimización del algoritmo adaptando el mismo a Giraph 1.2, implementación de búsqueda desde múltiples orígenes y múltiples destinos en un grafo.

Fecha de la presentación: Marzo del 2017