UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

TÍTULO: Estudio e implementación de una técnica de clustering dinámico para trabajar con flujos de datos.

AUTORES: Molina, Roberto Pedro.

DIRECTOR: Hasperué, Waldo.

CODIRECTOR:

ASESOR PROFESIONAL:

CARRERA: Licenciatura en Informática.

Resumen

El objetivo general de esta tesina es estudiar y analizar las técnicas y problemáticas existentes de clustering (agrupamiento) aplicadas sobre los flujos de datos, buscando técnicas que permitan un agrupamiento dinámico. También, se realizará una investigación y estudio sobre los frameworks o plataformas de procesamiento de flujos de datos actuales con el fin de analizar la viabilidad para generar técnicas de clustering sobre estos entornos. Tras los resultados de las investigaciones y estudios previos, se propone como objetivo particular para esta tesina, el desarrollo, implementación, evaluación y comparación de un algoritmo de clustering dinámico aplicado al tratamiento de flujos de datos.

Palabras Clave

Flujos de datos, DataStream, Cluster analysis, Data Streaming Clustering, Streaming processing, Apache Spark Streaming, Minería de Datos, Machine Learning, aprendizaje no supervisado, programación distribuida, Two-phase Learning, Silhouette, Time Window.

Conclusiones

Se presenta D3CAS, un algoritmo de clustering dinámico basado en densidad para el procesamiento de flujos de datos, por lo que el objetivo de diseñar e implementar una técnica de clustering dinámica se pudo lograr satisfactoriamente, brindando además, una solución capaz de trabajar en un entorno distribuido y escalable gracias a que la implementación fue llevada sobre el motor de Spark Streaming.
Los resultados obtenidos en los experimentos y comparaciones con otros algoritmos de clustering son alentadores, logrando una buena calidad en todos los datesets analizados.

Trabajos Realizados

Estudio del modelo de flujo de datos. Estudio de diferentes metodologías de tratamiento de flujos de datos. Investigación sobre el estado del arte de técnicas de Streaming Clustering. Análisis de diferentes enfoques y metodologías de Streaming Clustering. Investigación sobre Apache Spark y Apache Spark Streaming. Instalación, configuración y pruebas de conceptos sobre Spark Streaming. Investigación de procesos de debugging sobre Spark. Implementación de simuladores de flujos de datos a partir de datasets para realizar benchmark sobre tareas de clustering. Diseño e implementación de una técnica de clustering dinámico sobre flujos de datos (D3CAS). Evaluación de la calidad de los resultados obtenidos por D3CAS. Comparación de resultados entre D3CAS y Clustream.

Trabajos Futuros

Como trabajo futuro se propone la ejecución de D3CAS en un ambiente distribuido que permitan medir y mejorar el rendimiento del algoritmo presentado, con el fin de realizar comparaciones tanto a nivel de resultados como también a nivel de consumo de recursos como memoria, procesamiento, overhead de comunicación, consumo energético, etc. Otro aspecto a estudiar es el de ejecutar pruebas en un entorno real, consumiendo un flujo de datos real, como por ejemplo, se podría consumir los flujos de datos que brindan los servicios de Twitter o de redes de sensores.

Fecha de la presentación: 07-20118