Metaheurísticas en grandes volúmenes de datos combinados con *streaming* de datos en tiempo real

Ricardo Di Pasquale[1] Javier Marenco[2]

[1] Facultad de Ingeniería y Ciencias Agrarias, Pontificia Universidad Católica Argentina, Argentina
rdipasquale@uca.edu.ar
[2] Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina
jmarenco@dc.uba.ar

En los últimos años el procesamiento de corrientes (*streams*) de datos en tiempo real se ha incorporado definitivamente a los modelos de procesamiento *Big Data* existentes. Este tipo de procesamiento se da cuando uno o muchos emisores generan una corriente de datos en tiempo real de tal manera que si un receptor deja de "escuchar" un momento una de las corrientes de datos, la información que se omitió en esa ventana de tiempo se torna irrecuperable.

En este trabajo estamos interesados en estudiar las implicancias de incorporar el modelo *Big Data* de procesamiento de datos en metaheurísticas aplicadas a grandes volúmentes de datos estáticos.

Particularmente, se ha tomado una aplicación de descubrimiento de reglas en bases de datos (KDD) implementada mediante metaheurísticas distribuidas en plataforma *Apache Spark*. En dicha aplicación se busca descubrir reglas de asociación en una base de datos grande, por lo que la implementación distribuida se clasifica como un análisis de datos con estilo *Big Data*.

A la aplicación citada se le agrega la complejidad de procesar *streamings* de datos, que incorporan hechos a la base de datos de manera compatible con la información existente. Si, adicionalmente, la ponderación que se hace de la información más reciente (en tiempo real) es mayor a la valuación de la información histórica, se evidencia que la naturaleza del problema ha mutado. Muchas de las facilidades y pre-procesamientos posibles quedan invalidados por la incertidumbre generada por los datos a incorporarse.

Se presentarán los resultados obtenidos y las principales diferencias en los modelos de procesamiento. Se discutirá también si los problemas clásicos de optimización, o de *data mining*, o los que están en las fronteras pueden seguir siendo tratados de la misma manera al considerar el procesamiento de *streaming* en tiempo real, o si deben cambiarse los mecanismos de procesamiento de manera radical.