Tecnologías de Procesamiento de Datos Masivos

Ramiro Rivera, Luciano Bracco, Valentín Costa, Facundo Coto, Patricia Cristaldo, Lautaro Ramos, Natalia Rapesta, Juan Pablo Núñez, Soledad Retamar, Anabella De Battista
Grupo de Investigación en Bases de Datos, Departamento Ingeniería en Sistemas de Información, Fac. Reg. Concepción del Uruguay, Universidad Tecnológica Nacional
Entre Ríos, Argentina
{riverar, braccol, costav, cotof, cristaldop, ramosl, rapestan, nunezjp, retamars, debattistaa}@frcu.utn.edu.ar

Norma Edith Herrera
Departamento de Informática, Universidad Nacional de San Luis, San Luis, Argentina
nherrera@unsl.edu.ar

Resumen

Las grandes cantidades de datos que se producen en la actualidad, sumadas a su heterogeneidad y la velocidad con que se generan, hacen que las herramientas tradicionales de análisis de datos no resulten adecuadas para su recopilación, almacenamiento, gestión y análisis. En este contexto surge el término Big Data, en referencia a características como gran volumen, velocidad y variedad de producción de los datos, y a las herramientas que se utilizan para encontrar valor en los mismos. La posibilidad de hallar patrones y tendencias en estas grandes cantidades de datos impacta directamente en la toma de decisiones en áreas tan diversas como salud, genética, agro, predicciones climáticas, redes sociales, marketing, finanzas, educación, entre otras.
En este artículo se presentan los tópicos de interés del proyecto *Minería de Datos: su aplicación a repositorios de datos masivos*.
Palabras clave: Big Data, minería de datos, clustering, agrupamiento, streaming, gestión de proyectos.

Contexto

El presente trabajo se desarrolla en el ámbito del proyecto *Minería de Datos: su aplicación a repositorios de datos masivos (UTI3781TC)* del Grupo de Investigación en Bases de Datos, perteneciente al Departamento Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay.

1. Introducción

El término Big Data surge en referencia a conjuntos de datos cuyo tamaño supera la capacidad de procesamiento de las herramientas tradicionales de bases de datos. En general se habla de Big Data o Análisis de Big Data como sinónimos, ya que no sólo se desea hacer referencia a la gran cantidad y complejidad de los datos, sino también a las herramientas utilizadas para procesarlos y extraer conocimiento útil.
Algunas definiciones indican que Big Data puede definirse a partir de las siguientes características [1]:

- *Volumen*: órdenes superiores a Terabytes de datos.
- *Variedad*: distintos tipos de datos provenientes de diversas fuentes que pueden organizarse tanto en forma estructurada como no estructurada.
- *Velocidad*: referido a la velocidad de generación de los datos o a la rapidez con la que se generan y procesan los datos.
- *Variabilidad*: referido a la inconsistencia que puede presentar los datos en ocasiones, dificultando las tareas de análisis.
- *Valor*: gracias a la posibilidad de tomar decisiones al responder preguntas que antes no era posible, ofrece a la organización una ventaja estratégica.

Estos grandes repositorios de datos se generan desde fuentes tan diversas como redes sociales, instrumentos científicos, dispositivos móviles o redes de sensores, entre otros, y representan nuevos desafíos para su almacenamiento, tratamiento, distribución y análisis, ya que además de ser datos de gran volumen su complejidad es creciente. En este contexto cobra gran relevancia la posibilidad de hallar patrones en estos datos, pero más aún la posibilidad de explicar dichos patrones, ya que impacta directamente en la toma de decisiones aplicable en áreas tan diversas como salud, genética, agro, predicciones climáticas, redes sociales, marketing, finanzas, educación, entre otras. En este conexto, una actividad interesante es la detección de agrupamientos en repositorios de datos masivos y complejos.

Como se mencionó anteriormente, en la actualidad no es un obstáculo la capacidad de recopilar datos, en cambio sí lo es la capacidad de gestionar, analizar, sintetizar, visualizar y descubrir conocimiento en los datos recopilados de manera oportuna y en una forma escalable. Debido a que no es posible procesar estos datos tan masivos y complejos con herramientas tradicionales, han surgido nuevos algoritmos especialmente diseñados que aprovechan las características del procesamiento paralelo.

La Minería de Datos involucra e integra técnicas de diferentes disciplinas tales como tecnologías de bases de datos y data warehouse, estadística, aprendizaje de máquina, computación de alta performance, computación evolutiva, reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, y análisis de datos espaciales o temporales. En este proyecto se estudian procesos de Minería de Datos desde una perspectiva de bases de datos, con enfoque en técnicas eficientes y escalables a conjuntos de datos masivos.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de trabajo principal de nuestro proyecto de investigación es el estudio de técnicas de Minería de Datos aplicables a repositorios de datos masivos, atendiendo principalmente a su eficiencia y escalabilidad [2]. Los tópicos en los que se trabaja actualmente son: el tratamiento de datos en *streaming* y el análisis y comparación del funcionamiento de algoritmos de clustering y clasificación aplicables a datos masivos para posteriormente proponer mejoras a los algoritmos existentes o bien, nuevos algoritmos.

En la gestión de las actividades del proyecto se emplea la Metodología Fundacional para Ciencias de Datos, que consta de diez pasos y es algo similar a otras metodologías reconocidas para Minería de Datos, pero que enfatiza varias de las nuevas prácticas en ciencias de datos como el uso de grandes volúmenes de datos, la incorporación de análisis de texto dentro del modelo predictivo y la automatización de algunos procesos [3].

A continuación se detallan algunas de las actividades realizadas.

2.1.Análisis Bibliométrico

Se trabaja en análisis bibliométrico tradicional y alternativo, midiendo el impacto de publicaciones científicas en sus distintas modalidades de difusión. Actualmente se está elaborando un análisis cuantitativo de publicaciones de autores de instituciones argentinas en la bases de datos SCOPUS de Elsevier [4], accedida desde la Biblioteca Electrónica de Ciencia y Tecnología del Ministerio de Ciencia, Tecnología e Innovación Productiva de la Nación. En esta ocasión para las búsquedas se utilizó la palabra clave "pesticida". En algunos casos se aplicó como filtro que las publicaciones correspondiesen a Argentina para identificar y reunir los trabajos en los que al menos uno de los autores incluyera la mención de una institución argentina en los datos de afiliación institucional, a fin de poder comparar con la cantidad de publicaciones del resto del mundo. Los accesos a los datos se realizaron mediante de la API de Scopus [5] y mediante scripts desarrollados en R [6].

2.2. Algoritmos de clustering

El estudio de algoritmos de agrupamiento o *clustering* aplicables a datos masivos incluyó, en una primera etapa, la búsqueda bibliográfica sobre el funcionamiento y propósito general de esta clase de algoritmos y los diferentes tipos y características de los mismos. Posteriormente se analizaron las problemáticas que conlleva la masividad de los datos sobre los algoritmos de clustering, como por ejemplo la pérdida de eficiencia debido a la "Maldición de la Dimensionalidad". Se realizó una búsqueda sobre algoritmos desarrollados para sortear los problemas vinculados a los grandes conjuntos de datos, considerando tanto algoritmos específicos para la problemática como implementaciones de algoritmos tradicionales en plataformas de procesamiento de grandes cantidades de datos.
Luego del relevamiento se decidió trabajar con Halite [7], un algoritmo de agrupamiento novedoso diseñado para abordar las problemáticas de la alta dimensionalidad y el gran volumen de los conjuntos de datos. Se estudió en profundidad el funcionamiento del algoritmo, y se realizó una implementación del mismo a fin de poder comprobar su funcionamiento con conjuntos de datos de prueba del repositorio de la Universidad de California en Irvine (UCI, https://goo.gl/c8s2kg).
Actualmente se trabaja en una comparativa de Halite con algoritmos de clustering tradicionales [8] implementados en la herramienta Spark [9].

2.3. Algoritmos de clasificación

Debido a la gran cantidad de problemas de aplicación, la clasificación es uno de los temas más estudiados dentro del campo de la minería de datos. La tarea de clasificación consiste en la construcción de un modelo basado en instancias de datos clasificadas, que sea capaz de predecir o clasificar nuevos ejemplos de datos cuya etiqueta o clase se desconoce. La tarea de construir este modelo a partir de los denominados "datos de entrenamiento" no es una tarea sencilla debido a que éstos suelen ser ruidosos o a que algunas de las características más importantes son desconocidas, sumado a los costos que puede acarrear el cálculo de la similitud entre los datos de entrenamiento y el aprendizaje del modelo. Hemos abordado el estudio de los principales métodos de clasificación entre los que se pueden mencionar [9]:

- El *algoritmo de Naive Bayes* ha demostrado ser fácil de implementar, poseer eficiencia computacional y en la tasa de clasificación y obtener resultados precisos cuando el número de registros es grande.
- *Árboles de Decisión*: en esta técnica se divide el conjunto de datos de entrenamiento en dos o más partes utilizando alguno de sus atributos. Este proceso se repite hasta que solo haya datos de la misma clase en cada rama del árbol. La elección del atributo se puede llevar a cabo mediante el cálculo de la entropía o del índice Gini. Las principales ventajas de esta técnica se encuentran en que los modelos son fáciles de interpretar, tiene una implementación simple, soporta valores continuos y discretos y es tolerante al ruido presente en los datos. Una de las desventajas señaladas es que no tiene buen desempeño cuando el conjunto de datos de entrenamiento es pequeño o ruidoso.
- *Redes Neuronales*: entre los algoritmos más conocidos en esta área se encuentran los llamados Mapas Auto-organizados de Kohonen. Entre las principales ventajas se menciona la facilidad de uso, no necesita ser reprogramada y es aplicable a una amplia gama de problemas; mientras que las principales desventajas son la complejidad en determinar la cantidad de neuronas y capas necesarias, el tiempo de procesamiento elevado y que pueden tener un aprendizaje lento.
- *Algoritmos Genéticos*: estas técnicas inspiradas en la teoría de la evolución natural han sido aplicadas para la clasificación y debido a que realizan una

búsqueda global e independiente del dominio, los convierte en una herramienta robusta, escalable y aplicable en distintas etapas del proceso de extracción de conocimiento.

Adicionalmente, y debido a la evolución de las tecnologías de Big Data, un argumento que justifica su escalabilidad es que su funcionamiento facilita las implementaciones paralelas. Sin embargo una de las principales dificultades es el proceso de evaluación de individuos, aspecto en el que se continúa investigando.

2.4. Sistema de procesamiento de streaming de datos

Otra parte de los esfuerzos del grupo se hallan abocados al estudio del procesamiento de datos en streaming, tema que cobra cada vez más protagonismo, tanto a nivel académico como por su capacidad de aportar a la Inteligencia de Negocios de las organizaciones. Ya no es suficiente con ser capaces de procesar grandes cantidades de datos extraídos de repositorios o generados por las organizaciones, sino que deben ser procesados de manera rápida, o en "real time", además de generar información precisa. Los datos en streaming pueden provenir de diversas fuentes, como archivos de registros generados por los clientes que utilizan sus aplicaciones móviles o web, compras electrónicas, información de redes sociales, operaciones bursátiles o servicios geoespaciales. Algunos casos en lo que resulta útil el análisis de streaming de datos son la detección de fraudes, monitoreo de sistemas, intercambios comerciales y demás. El procesamiento de streams en tiempo real está diseñado para analizar y actuar en función de información a medida que la misma se genera, mediante el uso de consultas continuas (consultas del tipo SQL que operan sobre ventanas temporales e informacionales) [11]. Esto requiere un cambio de paradigma en cuanto al almacenamiento, obtención y procesamiento de la información. Las "bases de datos tradicionales" no fueron concebidas para este propósito por lo que debe hacerse uso de otras herramientas que otorguen la potencia y versatilidad que se requieren. Con este fin se busca la definición de una arquitectura capaz de procesar estos streams. Los componentes de la arquitectura deben ser capaces de interconectarse entre sí, proveer una alta tolerancia a fallas y permitir una escalabilidad elevada. A su vez, resultan atractivas aquellas herramientas basadas en el Software Libre, que se hallan respaldadas en el conocimiento colectivo de su comunidad. La propuesta en la que se trabaja es el desarrollo de la arquitectura de un sistema de procesamiento de datos en streaming, capaz de responder con una latencia máxima de 30 segundos a partir de un volumen de 100 eventos/seg.

En este desarrollo se utilizan las siguientes herramientas:

- *Kafka*: broker de mensajería, utilizado para centralizar la recepción de información sobre los eventos que se produzcan.
- *Zookeeper*: mecanismo de sincronización distribuido, que mantiene el estado y configuración de las demás piezas de software del sistema.
- *Docker*: tecnología de contenerización, distribución de aplicaciones y virtualización, su objetivo es garantizar sencillez de despliegue y posibilidad de escalado de la arquitectura.
- *Storm*: sistema distribuido de procesamiento de eventos en streaming, capaz de definir los "caminos" y "transformaciones" que sufren los eventos para poder extraer datos de interés para la organización.
- *Redis*: Base de datos NoSQL, del tipo clave-valor, utilizada para permitir la reconfiguración del sistema sin necesidad de Down-times.

3. Resultados obtenidos y esperados

Con este proyecto se espera proponer modificaciones o mejoras a los algoritmos de clustering o de clasificación existentes para datos masivos. Además se espera obtener una herramienta eficiente en el análisis de datos en streaming. Mediante la aplicación de la Metodología Fundacional para Ciencias de Datos se espera obtener resultados favorables en cada uno de los proyectos que se desarrollen.

4. Formación de Recursos Humanos

Este proyecto dio inicio a una nueva línea de investigación dentro del Grupo de investigación en Bases de Datos de la Fac. Reg. Concepción del Uruguay de la U.T.N.. Tres de los investigadores del proyecto están desarrollando tesis de maestría. En el proyecto colaboran dos becarios graduados con beca de iniciación a la investigación, que tienen previsto la realización de posgrados en el área temática del proyecto. Además participan en el proyecto cuatro becarios alumnos de la carrera Ingeniería en Sistemas de Información que inician su formación en la investigación. Tres de ellos están realizando su práctica supervisada en la temática de análisis de streaming.

5. Referencias

[1] Fan Wei and Bifet Albert. Mining big data: Current status, and forecast to the future. SIGKDD Explor. Newsl.,14(2):1–5, apr 2013.

[2] Larose Daniel T. Discovering Knowledge in Data: An Introduction to Data Mining. Wiley-Interscience, 2004.

[3] Foundational Methodology for Data Science. IBM. Disponible en http://goo.gl/w8tT9J . Accedido el 14 de Marzo de 2017.

[4] SCOPUS. http://www.scopus.com. Accedido 03/2017.

[5] Scopus API. http://http://goo.gl/umz8m6. Accedido 03/2017.

[6] R Project. https://www.r-project.org/ Accedido 03/2017.

[7] Robson L. F. Cordeiro, Christos Faloutsos, and Caetano Traina Junior. 2013. Data Mining in Large Sets of Complex Data. Springer Publishing Company, Incorporated.

[8] J.H. Orallo, M.J.R. Quintana, and C.F. Ramírez. Introducción a la minería de datos. Editorial Pearson, 2004. ISBN: 84 205 4091 9. 2004.

[9] Spark Streaming. http://spark-project.org Accedido 03/2017.

[10] Lyubchyk, L.,Grinberg, G. Real time recursive preference learning to rank from data stream. 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP) Data Stream Mining & Processing (DSMP), IEEE First International Conference on. :280-285 Aug, 2016

[11] Tyler Akidau. The world beyond batch: Streaming 101. https://goo.gl/xhPVZQ. Accedido 03/2017.