Minería de Datos y Visualización de Información

Esteban Schab, Ramiro Rivera, Luciano Bracco, Facundo Coto, Patricia Cristaldo, Lautaro Ramos, Natalia Rapesta, Juan Pablo Núñez, Soledad Retamar, Carlos Casanova, 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
{schabe, riverar, braccol, cotof, cristaldop, ramosl, rapestan, nunezjp, retamars, casanovac,debattistaa}@frcu.utn.edu.ar

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

Resumen

El procesamiento y análisis de las grandes cantidades de datos que se producen en la actualidad, posibilitan el hallazgo de patrones y tendencias ocultos en los mismos, que impacta directamente en la toma de decisiones en diversas áreas de estudios.
Se generan datos a gran velocidad y en grandes cantidades que requieren ser procesados para poder actuar de manera rápida. Como es el caso de la observación de turnos que se generan en entidades bancarias, donde hay momentos del día en que se requiere modificar los esquemas de atención, según la afluencia de determinadas categorías de clientes o el incremento de demandas de determinados servicios.
Existen numerosas técnicas de minería de datos aplicables a distintos casos de análisis de datos, que permiten obtener ventajas de esas grandes cantidades de datos almacenados.
En este artículo se presentan los tópicos de interés del proyecto *Minería de Datos: su aplicación a repositorios de datos masivos*, en el que se investigan tanto temas de minería de datos, como de visualización de información, como herramienta para representar de manera eficiente los resultados obtenidos.
Palabras clave: Minería de datos, streaming de datos, gestión de proyectos, visualización, scrapping.

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 Descubrimiento de Conocimiento en Bases de Datos consiste en el análisis automático exploratorio y modelado de grandes repositorios de datos e involucra áreas de conocimiento como inteligencia artificial, aprendizaje automático, estadística, sistemas de gestión de base de datos, técnicas de visualización de datos y medios que apoyan toma de decisiones.

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.
Como un subárea específica de la Minería de Datos se puede citar al Data Stream Mining,

que es el proceso de extraer conocimiento en estructuras de datos continuas y con rápidas transiciones [1]. Los data streams son datos que se generan de forma continua y a altas velocidades. El origen de dichos datos puede provenir de diversas fuentes, como registros generados por clientes que utilizan aplicaciones móviles, transacciones electrónicas, logs de navegación de una red de datos, información de redes sociales, datos provenientes de dispositivos wereables, entre muchos otros ejemplos.

El procesamiento de estos datos debe realizarse de forma secuencial y gradual registro por registro, o bien en ventanas de tiempo graduales. Los resultados de dicho procesamiento se utilizan para una amplia variedad de tipos de análisis, como correlaciones, agregaciones, filtrado y muestreo. Las conclusiones obtenidas a partir de dicho análisis aporta a las empresas visibilidad de numerosos aspectos del negocio y de las actividades de sus clientes, como la tasa de uso de un servicio, la actividad de un servidor, la ubicación geográfica de un móvil, personas o mercadería, la afluencia de determinado tipo de clientes, entre otros aspectos, y les permite responder con rapidez ante cualquier situación que surja. Por ejemplo, un banco podría analizar el incremento de determinada categoría de clientes en un momento dado y responder rápidamente habilitando más puestos de atención al cliente.

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 datos estructurados y no estructurados, atendiendo principalmente a su eficiencia y escalabilidad [2]. Actualmente se trabaja en la aplicación de técnicas de minería de datos en aplicaciones específicas como el procesamiento de streams de datos*, el análisis bibliométrico* y también en la temática *Agenda setting*.

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 [3], accedida desde la Biblioteca Electrónica de Ciencia y Tecnología del Ministerio de Ciencia, Tecnología e Innovación Productiva de la Nación y a través de la API proporcionada por SCOPUS [4], utilizando scripts desarrollados en lenguaje R [5]. Para las búsquedas se establece una palabra o frase clave. 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.

Se comenzó trabajando en índices tradicionales de análisis de publicaciones (conocido como Modo 1), que se basan principalmente en analizar las publicaciones realizadas en revistas con referato pagas revistas pagas con referato. En la actualidad, se está trabajando en la actualidad en el denominado Modo 2, estudiando instituciones de pertenencia de los artículos provenientes de Scopus y además de en fuentes como Altmetric, en este último caso se busca información publicada en blogs de ONGs, institucionales, etc. Está previsto realizar un análisis del impacto de publicaciones en redes sociales como Facebook o Twitter.

2.2. Agenda setting

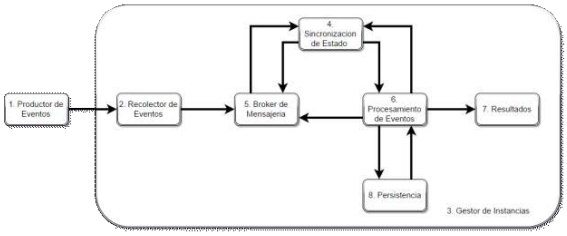
El término Agenda Setting hace referencia a la influencia que tienen los medios de comunicación en la fijación de temas en la opinión pública [6].

Se comenzó a realizar un trabajo de medición de los efectos de la instalación de asuntos en la agenda pública tomando como base artículos escritos sobre diferentes temáticas en medios digitales de relevancia para

determinar los tópicos que tratan y luego analizar su difusión en redes sociales empleando técnicas de minería de textos y procesamiento de lenguaje natural [7].

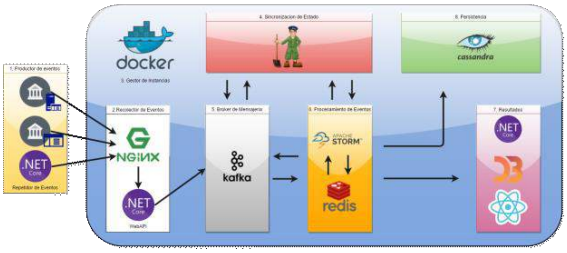
2.3. Sistema de procesamiento de streaming de datos

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 de aplicación del 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) [8]. 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 en este esquema de aplicaciones.
Con este fin se trabajó en la definición de una arquitectura capaz de procesar streams de datos. Se propuso una arquitectura genérica, la cual representa los componentes necesarios para realizar la captura y procesamiento de los streams y sus interacciones.



Los componentes de la arquitectura deben ser capaces de interconectarse entre sí, proveer una alta tolerancia a fallas y permitir una escalabilidad elevada. En este esquema se pensó en la utilización de herramientas basadas en Software Libre, que se hallan respaldadas en el conocimiento colectivo de su comunidad. La arquitectura propuesta permitirá realizar el procesamiento de datos en streaming y será capaz de responder con una latencia máxima de 30 segundos a partir de un volumen de 100 eventos/seg.
A partir de la arquitectura genérica propuesta y los requerimientos antes mencionados, se planteó una Arquitectura para el Procesamiento de Streams de Datos que se encuentra en desarrollo haciendo uso de 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, con el fin de 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.



2.4. Transferencia tecnológica a industrias de la zona

En el marco de un convenio con una empresa local de desarrollo de software, se está realizando el desarrollo del prototipo para el procesamiento y posterior análisis de streaming de datos provenientes de las aplicaciones que esta empresa comercializa en bancos, con el objetivo de ofrecer información agregada para la toma de decisiones y que pueda retroalimentar dicha aplicación para automatizar ciertas decisiones a futuro.

En conjunto con el Grupo de Investigación y Desarrollo en Innovación y Competitividad del Departamento Licenciatura en Organización Industrial, de la FRCU-UTN, se encuentra en desarrollo el proyecto "Fortalecimiento de la Gestión productiva integral en PyMEs del sector metalmecánico del Parque Industrial de Concepción del Uruguay, Entre Ríos". Dicho proyecto ha resultado seleccionado en el marco de la convocatoria "Agregando Valor" (edición 2017) de la Secretaría de Políticas Universitarias de la Nación, orientada a la presentación de proyectos de vinculación tecnológica de alto impacto con la finalidad de transferir conocimientos y tecnologías innovadoras al sector socio-productivo nacional.

2.5. Visualización de datos

La generación y el almacenamiento de grandes volúmenes de información hacen que el mismo pase desapercibido y muchas veces se pierde la oportunidad de encontrar valor en ella. La visualización de datos es el proceso de representación de datos, en formato gráfico, de una manera clara y eficaz. Se convierte en una herramienta poderosa para el análisis e interpretación de datos grandes y complejos, volviéndose un medio eficiente en la transmisión de conceptos en un formato universal [9, 10].

En este proyecto se trabaja en el análisis de técnicas y herramientas de visualización de datos, para mejorar los procesos de comunicación de resultados de las actividades que desarrolla el grupo. A partir de la generación de visualizaciones de dichos resultados, se logra una mejor comprensión de los datos. Entre las herramientas utilizadas actualmente se encuentran Tableau[11], Gephi[12], D3js[13], React D3[14] y Shiny[15].

2.6. Aplicación de metodología para la gestión de proyectos de Minería de Datos

En la gestión de las actividades de cada una de las líneas de investigación y desarrollo del proyecto se emplean fundamentos de metodologías ágiles. [16, 17] Partiendo de la propuesta metodológica de CRISP-DM [18] se realizó una adaptación empleando dichos fundamentos ágiles. Se espera poder formalizar dicha adaptación de CRISP-DM como una propuesta de metodología ágil para la gestión de proyectos de ciencia de datos.

3. Resultados obtenidos y esperados

Con este proyecto se espera lograr aplicaciones novedosas de técnicas y herramientas de minería de datos, en particular en áreas de estudio como bibliometría, la teoría de establecimiento de agenda. Además se espera obtener una herramienta eficiente en el análisis de datos en streaming. Todas estas iniciativas se desarrollan mediante la aplicación de una metodología ágil para proyectos de ciencias de datos.

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..

Dos 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, dos de ellos han realizado su Práctica Supervisada en el marco del proyecto.

5. Referencias

[1] Khan, Latifur, and Wei Fan. 2012. "Tutorial: Data Stream Mining and Its Applications." In Database Systems for Advanced Applications, eds. Sang-goo Lee et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 328–329.

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

[3] SCOPUS. http://www.scopus.com Accedido 03/2018.

[4] Scopus API. https://goo.gl/mqpFpA Accedido 03/2017.

[5] R Project. https://www.r-project.org/ Accedido 03/2018.

[6] M. McCombs and D. Shaw. The agenda-setting function of mass media. Public opinion quarterly, 36(2):176–187, 1972.

[7] Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. A computational analysis of agenda setting. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion), 323-324, 2014.

[8] Tyler Akidau. The world beyond batch: Streaming 101. https://goo.gl/xhPVZQ. Accedido 03/2018.

[9] Sadiku, Matthew (2016). Data Visualization. International Journal of Engineering Research And Advanced Technology(IJERAT).Volume. 02. Issue.12. p. 11-16.

[10] Finch, Jannette L., and Angela R. Flenner. 2017. "Using Data Visualization to Examine an Academic Library Collection." College & Research Libraries 77(6). https://goo.gl/fAeW3w (March 18, 2018).

[11] Tableau. https://www.tableau.com/es-es Accedido 03/2018.

[12] Gephi. https://gephi.org/ Accedido 03/2018.

[13] Data-Driven Documents. https://d3js.org/ Accedido 03/2018.

[14] React D3. http://www.reactd3.org/ Accedido 03/2018.

[15] Shiny from R Studio. https://shiny.rstudio.com/ Accedido 03/2018.

[16] Ken Schwaber and Jeff Sutherland. The scrum guide. Scrum Alliance, 2011, vol. 21.

[17] Manifesto for Agile Software Development. Agile Alliance. https://goo.gl/xRFCVL . Accedido 03/2018.

[18] Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth. CRISP-DM 1.0 Step-by-step data mining guide. 2000.