VISUALIZACIÓN EN CIENCIA DE DATOS

Franco Castro, Mag. Graciela Elida Beguerí, Mag. María Alejandra Malberti

Instituto de Informática / Departamento de Informática /Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan

Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, Teléfonos: 4260353, 4260355 Fax 0264-4234980, Sitio Web: http://www.exactas.unsj.edu.ar
poyuyo2013, grabeda , amalberti @gmail.com

RESUMEN

El vertiginoso aumento de datos generados en los últimos años, ha servido de incentivo al desarrollo y evolución de la Ciencia de Datos. Big Data es un término aplicado a conjuntos de datos cuyo tamaño o tipo está más allá de la capacidad de las bases de datos relacionales tradicionales tanto para capturar, gestionar o procesar los datos con baja latencia. Esos datos provienen de sensores, video/audio, redes, archivos de registro, transacciones, web y redes sociales, gran parte de ellos generados en tiempo real y en gran escala. El análisis de Big Data permite a diferentes tipos de usuarios (analistas, investigadores, usuarios comerciales) tomar decisiones utilizando los datos que antes eran inaccesibles o inutilizables. Mediante el uso de técnicas avanzadas de análisis como análisis de texto, aprendizaje automático, análisis predictivo, minería de datos y estadísticas, las organizaciones pueden analizar diversas fuentes de datos no tratadas previamente para obtener nuevas ideas que les permitan tomar mejores y más rápidas decisiones. A las cuatro V, que representan las dimensiones de Big Data propuestas por IBM: Volumen, Variedad, Veracidad y Velocidad, se le suma una quinta V, o dimensión: Visualización, que hace referencia a la representación visual, comprensible de los datos. En el marco de Ciencia de Datos, esta línea de investigación propone analizar y caracterizar diferentes estrategias y herramientas de búsqueda de conocimiento para la toma de decisiones, según sus potencialidades de Visualización de Información y principios de Deep Learning. Éstas se aplicarán a conjuntos de datos obtenidos desde diversas fuentes, en especial los disponibles bajo el nombre Open Data. De acuerdo a la naturaleza y magnitud de los datos, se considerarán variadas herramientas de software libre disponibles en el mercado, atendiendo a las potencialidades de visualización que las mismas ofrecen.

Palabras clave: Visualización, Visualización de Datos, Visualización de Información, Ciencia de Datos

1. CONTEXTO

La línea de investigación está inserta en el proyecto bianual 2018-2020 “Visualización y Deep Learning en Ciencia de Datos” que se desarrollará en el ámbito de la FCEFN-UNSJ.
En el contexto de la misma se pretende dar continuidad a los proyectos “Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuario”, “Búsqueda de Conocimientos en Datos Masivos” y “La Ciencia de Datos en grandes colecciones de datos” llevados adelante a partir de 2012.
Considerando al usuario como el destinatario del proceso de búsqueda de conocimiento en datos, se investigará sobre aspectos de interpretación y percepción, con el propósito de clasificar las herramientas tratadas, en el marco de Visualización de Información y niveles de usuarios.

2. INTRODUCCIÓN

En la reunión patrocinada por la Division of Advanced Scientific Computing - National Science Foundation – EEUU en octubre de 1986, surgió la iniciativa de considerar a la

Visualización, como un método informático que "transforma lo simbólico en geométrico", y que estudia además los mecanismos que permitan percibir, usar y comunicar la información visual.

Existen diversas perspectivas desde las cuales la visualización es abordada, entre las que se destacan la Perspectiva Cognitiva, que se apoya en la Psicología Cognitiva para analizar el sistema humano de la visión (percepción y procesamiento), y justificar las posibilidades de la visualización para ampliar el entendimiento; la Perspectiva Tecnológica, que incluye un amplio conjunto de técnicas o métodos provenientes de otras disciplinas, tales como la estadística, la minería de datos, y el procesamiento de imágenes, para facilitar el análisis desde aspectos cuantitativos y cualitativos y la Perspectiva Comunicacional, que considera la visualización como una ayuda eficaz para comunicar ideas.

Así como se distinguen los términos Dato, Información y Conocimiento, diversos autores proponen áreas de conocimiento o subcampos de visualización, entre los que se encuentran Visualización de Datos y Visualización de Información.

Entre las definiciones de Visualización de datos se encuentra la propuesta por Friendly y Denis (2006), que la consideran como "la ciencia de la representación visual de los datos", prestando especial atención a los gráficos estadísticos. En general este tipo de visualización se orienta a la representación de los datos con fines exploratorios, encontrándose entre los gráficos habitualmente usados Tablas, Diagramas de Cajas y Bigotes, Gráficos de Barra, Gráficos de Línea, Gráficos Circulares, Gráficos de Dispersión, Gráficos de Burbujas, Infografías y Nubes de Palabras.

Respecto a Visualización de Información, Ltifi y otros. (2009) la definen como "Explotar las aptitudes perceptivas naturales del usuario para comprender cualitativamente los volúmenes de información". Por su parte Ignasi Alcalde, en su artículo Visualización de información: ¿arte o ciencia? la considera como "… la representación y presentación de datos que explota nuestra capacidad de percepción visual con el fin de ampliar el conocimiento".

El proceso de visualización de información presentado en la figura 1 involucra la visualización de datos en sus tres primeras etapas: Datos-Información y Representación Gráfica. La cuarta etapa considera la comprensión a alcanzar, por parte del usuario, de la representación, involucrando aspectos de percepción e interpretación.

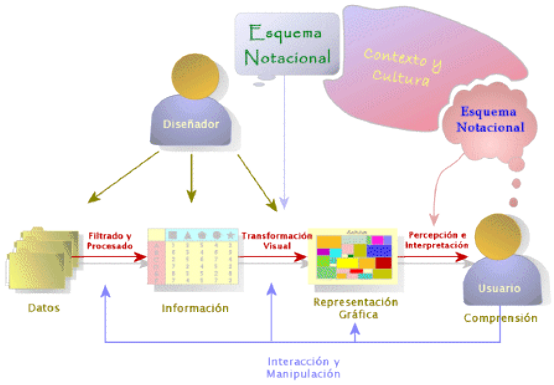

Figura 1 – Diagrama de Infovis http://www.infovis.net/printMag.php?num=187&lang=1

En este marco, diversos estudios sobre la percepción y cognición humana vinculados con la Visualización de Información se han realizado en los últimos años.

3. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En el marco del proyecto, en lo que refiere a visualización, se pretende analizar y caracterizar diferentes estrategias de búsqueda de conocimiento en datos.

También se procura:

- Evaluar las ofertas de software libre apropiadas al área Ciencia de Datos.
- Estudiar y analizar diferentes conjuntos de datos masivos a procesar. Evaluar herramientas de software libre para arquitecturas secuenciales, paralelas y distribuidas.
- Analizar las distintas alternativas de representación de datos de entrada y determinar cuáles favorecen al desempeño de los algoritmos destinados a la búsqueda de conocimiento en datos.
- Investigar sobre aspectos de interpretación y percepción relacionados con mecanismos de visualización de información, en el contexto de búsqueda de conocimiento en datos.
- Proponer una clasificación de herramientas disponibles, en el marco de visualización de información y niveles de usuarios, y
- Caracterizar a los usuarios de acuerdo a las potencialidades de las herramientas analizadas.

4. RESULTADOS OBTENIDOS Y ESPERADOS

Atendiendo a la disponibilidad de diversas herramientas de visualización, es de interés considerar sus principales características desde el punto de vista del usuario así como capacidad de transmitir resultados surgidos de un proceso de búsqueda de conocimiento en datos.

Otro aspecto a considerar son los tipos de datos que las herramientas pueden tratar. A la hora de considerar datos, una buena opción son los datos abiertos (Open Data), esto es una filosofía y una práctica que persigue que determinados tipos de datos estén disponibles de forma libre, sin restricciones de derechos de autor, de patentes o de otros mecanismos de control (Open Knowledge Foundation). Estos conjuntos de datos se encuentran almacenados en distintos formatos. El primer desafío que enfrentamos es reconocer los formatos de uso más general y tratarlos con herramientas de software libre, en este caso Knime, en las distintas etapas de Ciencia de Datos. El tipo de archivo depende en gran parte del tipo de dato que contiene, por ejemplo para representar datos georeferenciados, alternativas muy usadas son los KML o Shapefile. Para otro tipo de datos los formatos xls, csv, xml y json son los más utilizados. Una vez que se logra leer los distintos formatos se procede a analizar, clasificar y filtrar los datos contenidos según su tipo.

En la figura 2 se observa un ejemplo en Knime, en el cual se realiza la lectura de un archivo xml. Una vez leído el archivo es procesado con el nodo XPath, que permite separar los atributos del archivo xml en diferentes columnas.

Figura 2 – Workflow de lectura y procesamiento de un archivo xml.

Finalmente la información obtenida es visualizada en un histograma como se muestra en la figura 3, respondiendo a lo que se considera Visualización de Datos.

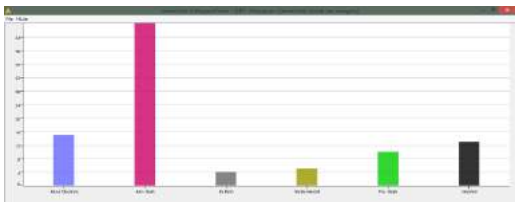

Figura 3 – Histograma representando la columna category

5. FORMACIÓN DE RECURSOS HUMANOS

La ejecución de las tareas proyectadas incidirá en una formación más profunda de los integrantes del equipo de investigación, en las recientes tecnologías de la Ciencia de Datos, en particular lo relativo a Visualización y Deep Learning. Este aspecto beneficiará de manera directa a las carreras del Departamento de Informática- UNSJ, pues las temáticas abordadas están vinculadas con las

materias en las cuales se desempeñan los integrantes de este proyecto.
Se dará continuidad a dos Becas de Investigación de Alumnos Avanzados, UNSJ, en los temas de paralelización de algoritmos y visualización, así como a dos trabajos finales de grado aplicados a diferentes conjuntos de datos y algoritmos de Deep Learning,
También se prevé la generación de nuevos trabajos finales de grado para las carreras del DI, y de tesis de maestría.

6. BIBLIOGRAFÍA

- Alcalde, Ignasi. *Visualización de información: ¿arte o ciencia?* Recuperado 30/11/2017 https://ignasialcalde.es/visualizacion-de-informacion-arte-o-ciencia/
- Buduma, N., & Locascio, N. (2017). *Fundamentals of Deep Learning: Designing Next-generation Machine Intelligence Algorithms.* " O'Reilly Media, Inc."
- Iliinsky, N., & Steele, J. (2011). *Designing data visualizations: representing informational relationships.* " O'Reilly Media, Inc.".
- Fayyad, UM, Wierse, A., y Grinstein, GG (Eds.). (2002). *Visualización de información en minería de datos y descubrimiento de conocimiento*. Morgan Kaufmann.
- Friendly, M. & Denis, D. (2006). *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* Recuperado 22/11/2017- http://web.calstatela.edu/curvebank/index/milestone.pdf
- Karimi, H. A. (Ed.). (2014). *Big data: Techniques and technologies in geoinformatics*. CRC Press.
- Ltifi, H., Ayed, M. B., Alimi, A. M., & Lepreux, S. (2009, May). *Survey of information visualization techniques for exploitation in KDD.* In Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on (pp. 218-225). IEEE.
- **McCandless, D.** (**2010**). *The beauty of data visualization.* John Wiley & Sons, TED website.
- Yuk, M., & Diamond, S. (2014). *Data visualization for dummies.* John Wiley & Sons.
- KNIME Analytics Platform. https://www.knime.com/knime-analytics-platform