Análisis visual de datos multidimensionales

Ganuza, M. Luján[1,2,3] , Antonini, Antonella S.[1,2,3], Luque, Leandro E.[1,3], Selzer, Matías[1,2,3] Larrea, Martín L.[1,2,3] , Tanzola, Juan E.[4,5], Asiaín, Lucía[4,5], Ferracutti, Gabriela R.[4,5], Gargiulo, M. Florencia[4,5] , Bjerg, Ernesto A.[4,5], Castro, Silvia M.[1,2,3]

[1]Laboratorio de I+D en Visualización y Computación Gráfica (VyGLab) (UNS-CIC)

[2]Dpto. de Cs. e Ing. de la Computación, Universidad Nacional del Sur (DCIC-UNS)

{mlg, antonella.antonini, leandro.luque, matias.selzer, mll, smc}@cs.uns.edu.ar

[3]ICIC, Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET)

[4]INGEOSUR, Instituto de Geología (UNS-CONICET)

[5]Dpto. de Geología, Universidad Nacional del Sur (DG-UNS)

{jetanzola,lasiain,ebjerg}@ingeosur-conicet.gob.ar, {gferrac,mfgargiulo}@uns.edu.ar

RESUMEN

La representación visual de datos altamente dimensionales facilita la comprensión y el análisis de las complejas relaciones entre múltiples características en un espacio multidimensional. A medida que aumenta la dimensionalidad de los datos, la visualización se vuelve más desafiante, ya que los espacios multidimensionales son difíciles de comprender y su representación visual requiere considerar numerosas variables y sus interrelaciones. Existen métodos de visualización para datos multidimensionales, pero aún enfrentan desafíos como la pérdida de información y la oclusión. Se necesita un progreso en la creación de métodos de descubrimiento visual más escalables y efectivos. En este contexto, nos enfocamos en mejorar las técnicas de visualización para datos multidimensionales con y sin pérdida de información. Además, proponemos la evaluación de las técnicas propuestas mediante métodos de seguimiento ocular.

Palabras claves: Análisis Visual de Datos Multidimensionales, Visualización de Datos, Visualización sin Pérdida de Información.

CONTEXTO

Este trabajo se realiza en estrecha colaboración con investigadores del INGEOSUR-CONICET (Instituto Geológico del Sur), del Departamento de Geología de la Universidad Nacional del Sur, y el VyGLab (Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica) del Departamento de Ciencias e Ingeniería de la Computación (DCIC-UNS), Instituciones de investigación de reconocido prestigio tanto nacional como internacional.

1. INTRODUCCIÓN

La visualización es una herramienta ampliamente utilizada en diversas áreas de las ciencias en las que se generan volúmenes de datos cada vez más grandes y difíciles de analizar y comprender sin un soporte visual adecuado [Cle93, CMS99].

Al aumentar la dimensionalidad de los datos, lograr una visualización expresiva y efectiva de éstos se convierte en una tarea cada vez más compleja, ya que los espacios multidimensionales, debido a su naturaleza, son difíciles de entender. Dado que sólo los datos 2-D y 3-D pueden ser visualizados directamente en el mundo físico 3-D, la visualización de los datos n-D se hace cada vez más difícil a medida que aumenta la cantidad de dimensiones de los mismos. Su representación visual exige considerar simultáneamente una gran cantidad de variables y sus interrelaciones.

Por ende, necesitamos herramientas de visualización mejoradas para representar n-

dimensiones en 2-D o 3-D sin tener pérdida de información. Muchas de las técnicas tradicionales sólo producen una visión parcial de los datos, ya sea por su dimensionalidad y/o su cantidad, lo que dificulta considerablemente detectar posibles relaciones que pudieran existir entre las variables que intervienen en un proceso. Esto genera la imposibilidad de tratar el fenómeno en toda su dimensionalidad y/o su cantidad.

Se han desarrollado diversos métodos de visualización para datos multidimensionales y se han utilizado con éxito para llevar a cabo diversas tareas sobre conjuntos de datos particulares. Sin embargo, tanto la pérdida de información como la oclusión siguen siendo un reto en las visualizaciones de datos n-D. En este contexto, se ha detectado la necesidad de ampliar la clase de representaciones visuales para datos multidimensionales en general, y de representaciones sin pérdida de información en particular. Estos métodos tienen ventajas como potenciadores de las capacidades cognitivas humanas en tareas de descubrimiento de patrones en datos multidimensionales [Kov18].

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El objetivo general de este trabajo se centra en contribuir al desarrollo de tecnologías y soluciones en torno al análisis visual de datos multidimensionales y su evaluación. En este contexto se plantean dos líneas principales de investigación:

El diseño y desarrollo de técnicas y herramientas para la visualización de datos multidimensionales

Los principales desafíos en la visualización de datos multidimensionales son la oclusión, la pérdida de información n-dimensional significativa al representar datos multidimensionales en 2-D o 3-D y las dificultades para encontrar representaciones visuales efectivas y expresivas de los datos. Uno de los abordajes desarrollados para manejar la multidimensionalidad de los datos, corresponde a la reducción de dimensionalidad, que consiste en el proceso de reducción de la cantidad de dimensiones que se trata, es decir que cada punto de datos n-D se proyecta en un solo punto 2-D o 3-D. Si bien la reducción de dimensiones es una de las técnicas de abstracción de datos más utilizadas en analítica visual, su utilización produce pérdida de información que, en muchos casos, es irreversible. En general, no hay manera de restaurar completamente los puntos n-D de estos dos o tres (2-D o 3-D) componentes excepto en algunos conjuntos de datos muy particulares. Las técnicas tradicionales de visualización, tales como scatterplots, histogramas, etc., por ejemplo, sólo producen una visión parcial de la información. Tales métodos no representan los datos multidimensionales completamente ni permiten su completa restauración desde su representación 2-D o 3-D.

Los enfoques disponibles actualmente para superar estas limitaciones son bastante limitados. Las coordenadas paralelas y las coordenadas radiales/en estrella son, hoy en día, los métodos de visualización de datos multidimensionales reversibles y sin pérdidas más utilizados; sin embargo, presentan varios inconvenientes, siendo la oclusión uno de los más significativos.

Una ventaja relevante de las visualizaciones sin pérdida de información es que un analista puede comparar una mayor cantidad de atributos de datos (potencialmente todos) que en las visualizaciones con pérdidas. Sin embargo, el diseño y desarrollo de nuevos métodos sin pérdida de información es escaso. En este contexto, se propone avanzar en el conocimiento de técnicas de

visualización para datos multidimensionales con y sin pérdida de información. Dada la amplia variedad de tipos de datos existentes nos centraremos en datos multidimensionales y multivariados temporales, espaciales y espacio-temporales que son comunes a diferentes dominios. En particular, se validarán las soluciones propuestas mediante la utilización de conjuntos de datos provenientes de las ciencias geológicas dado que el trabajo de laboratorio de los geólogos constituye un área de aplicación muy prometedora para la visualización de datos multidimensionales.

Evaluación de las nuevas técnicas diseñadas utilizando dispositivos de seguimiento ocular

La evaluación de las visualizaciones es fundamental para el avance de las técnicas existentes, ya que solo a través de la evaluación podemos demostrar la efectividad de una técnica, identificar posibles deficiencias y diseñar mejoras adicionales para respaldar y amplificar la cognición de manera más efectiva [FNS17].

A lo largo de los años, se han desarrollado diversos enfoques de evaluación para diferentes tipos y técnicas de visualización. Muchos de éstos registran el rendimiento del usuario en términos de tiempos de respuesta y tasas de error, así como algunos comentarios cualitativos de los usuarios [Duc17].

El eye tracking (seguimiento ocular), por otro lado, es una técnica avanzada que registra el comportamiento espacio temporal del movimiento ocular proporcionando variables adicionales más allá de las medidas estándar como los tiempos de finalización y las tasas de error. El participante puede concentrarse completamente en la solución de la tarea mientras inspecciona el estímulo y, al mismo tiempo, el dispositivo de seguimiento ocular está registrando la atención visual en forma de puntos de fijación con sus duraciones de fijación. Estos datos contienen información más detallada sobre el comportamiento del usuario y también plantean nuevos desafíos, dado que requieren tecnologías más avanzadas para registrar los datos, análisis algorítmicos más complejos y visualizaciones interactivas para analizarlos.

En comparación con los enfoques tradicionales de evaluación, que miden indirectamente las visualizaciones mediante entrevistas y experimentos controlados, el seguimiento ocular proporciona información directa sobre los movimientos oculares de los participantes durante un experimento, convirtiéndolo en un método apropiado para evaluar visualizaciones.

3.RESULTADOS OBTENIDOS Y ESPERADOS

De las líneas de trabajo delineadas se han obtenido resultados parciales. Los miembros del equipo de investigación trabajan desde hace tiempo y de manera sostenida en el diseño y desarrollo de soluciones en torno a datos multidimensionales para distintos dominios de aplicación [SBG+18; LGA+21; SUS+21; AGC+22; ALG+23; KSB+22], y aplicado a ciencias geológicas en particular [GCF+12; GFG+14; GGF+15; GFG+17; AGF+21; AFG+23; ALG+23]. Adicionalmente, en colaboración con el Laboratorio de Desarrollo en Neurociencias Cognitivas (LDNC) del Departamento de Ingeniería Eléctrica y Computadoras (DIEC) perteneciente a la Universidad Nacional del Sur (UNS), se trabaja en la utilización de técnicas de seguimiento ocular y en el diseño y desarrollo de herramientas de visualización para datos provenientes de dispositivos de seguimiento ocular [LGC+21; LGC+22]. Finalmente, y con respecto a la evaluación de las técnicas propuestas, se participa activamente en el ámbito de la Verificación y

Validación del Software aplicado a técnicas de visualización en particular [LSG+22; LUG23].

Las investigaciones que se vienen llevando a cabo en el grupo de trabajo tienen interesantes usos en la búsqueda, evaluación y exploración de recursos naturales renovables y no renovables, emplazamientos de obras civiles e identificación y remediación de riesgos ambientales y uso de suelos, entre muchas otras actividades. Incorporar en los análisis de los datos cuidadosas observaciones y medidas de campo, características microscópicas y macroscópicas de muestras de rocas y resultados de análisis geoquímicos e isotópicos permite vislumbrar la gran utilidad de la visualización de datos multidimensionales para el geólogo en su trabajo de laboratorio. De este modo, éste podrá contar con herramientas informáticas que le brinden el soporte necesario para analizar sus datos. Estas nuevas tecnologías contribuirán efectivamente a lograr una mejor comprensión de la interacción entre los procesos geológicos y las composiciones de minerales, rocas, sedimentos, fluidos, emanaciones de gases, entre otros, en la configuración geológica de una región o localidad en particular y/o en la contaminación/remediación ambiental de algún sector de interés.

Debe señalarse que las tareas a desarrollar para alcanzar los objetivos planteados se llevarán a cabo en estrecha colaboración tanto con integrantes del INGEOSUR, como con actores del ámbito nacional e internacional.

4. FORMACION DE RECURSOS HUMANOS

A continuación, se detallan proyectos de investigación, tesis finalizadas y en desarrollo y becas obtenidas, dedicadas a temáticas vinculadas con las líneas de investigación presentadas.

Proyectos de Investigación

-PIBAA - CONICET (2872021010 0824CO) *Análisis Visual de Datos Multidimensionales sin Pérdida de Información*. Directora: Dra. M. Luján Ganuza.

-PGI 24/ZN38 "Tecnologías Inmersivas y Visualización Situada aplicadas a Geociencias". Directora: Dra. M. Luján Ganuza.

-PGI 24/N050 *Verificación y Validación de Representaciones Visuales y sus Interacciones.* Director: Dr. Martín L. Larrea.

Tesis en Desarrollo

-*Análisis Visual de Datos Multidimensionales*, Doctorado en Cs. de la Computación. Antonella S. Antonini. Directora: Dra. Silvia Castro. Codirectora: Dra. M. Luján Ganuza.

-*Análisis visual de datos provenientes de registradores oculares*, Doctorado en Cs. de la Computación. Leandro Luque. Directoras: Dra. Silvia Castro y Dra. M. Luján Ganuza.

Becas

-Antonella S. Antonini. *Análisis Visual de Datos Multidimensionales.* Beca doctoral 2018 CONICET. Adjudicada a partir de abril de 2019. Directores: Dra. Silvia Castro - Dr. Ernesto Bjerg.

-Leandro Luque. *Análisis visual de datos provenientes de registradores oculares.* Beca doctoral 2018 CONICET. Adjudicada a partir de abril de 2019. Directores: Dra. Silvia Castro - Dr. Osvaldo Agamennoni.

5. BIBLIOGRAFÍA

[AFG+23] Antonini, A., Ferracutti, G., Gargiulo, F., Bjerg, E., Castro, S. & Ganuza, M. L. Análisis visual de datos multidimensionales de química mineral correspondientes a oxiespinelos de xenolitos del manto con Spinel Web. Caso de estudio. En Actas del 14° Congreso de Mineralogía, Petrología Ígnea y Metamórfica, y Metalogénesis - 14° MinMet y 5° PIMMA - Serie D, Publicación Especial 16 (pp. 36-42).

[AGC+22] Antonini, A. S., Ganuza, M. L. and Castro, S. M. VISUEL - A Web Dynamic Dashboard for Data Visualization. J. Comput. Sci. Technol. 2022, 22(1), 42-57.

[AGF+21] Antonini, A. S., Ganuza, M. L., Ferracutti, G., Gargiulo, M. F., Matković, K., Gröller, E., …Castro, S. M. Spinel web: an interactive web application for visualizing the chemical composition of spinel group minerals. Earth Sci. Informatics 2021, 14(1), 521-528.

[ALG+23] Antonini, A. S., Luque, L., Ganuza, M. L. & Castro, S. M. Towards a Taxonomy for Non-Paired General Line Coordinates: A Comprehensive Survey. Int. J. Data Sci. Anal. 2023, (15), 133-158.

[Cle93] Cleveland, W. S. Visualizing Data. Hobart Press, 1993.

[CMS99] Card, S., Mackinlay, J., Shneiderman, B., Readings in Information Visualization – Using Vision to Think, Morgan Kaufmann, 1999.

[Duc17] Duchowski, T. Eye tracking: methodology theory and practice. Springer, 2017.

[FNS17] Fu, B., Noy, N. F., and Storey, M.-A. Eye tracking the user experience–an evaluation of ontology visualization techniques. Semantic Web 8, 1 2017, 23–41.

[GCF+12] Ganuza, M. L., Castro, S. M., Ferracutti, G., Bjerg, E. A., and Martig, S. Spinelviz: An interactive 3d application for visualizing spinel group minerals. Comput. Geosci. 48 (2012), 50–56.

[GFG+14] Ganuza, M. L., Ferracutti, G., Gargiulo, M. F., Castro, S. M., Bjerg, E. A., Gröller, E., and Matkovic, K. The spinel explorer - interactive visual analysis of spinel group minerals. IEEE Trans. Vis. Comput. Graph. 20, 12 (2014), 1913–1922.

[GFG+17] Ganuza, M.L., Ferracutti, G., Gargiulo, M. F., Castro, S. M., Bjerg, E. A., Gröller, E., and Matkovic, K. Interactive visual categorization of spinel-group minerals. SCCG 2017.

[GGF+15] Ganuza, M.L., Gargiulo, M.F., Ferracutti, G., Castro, S. M., Bjerg, E. A., Gröller, E., and Matkovic, K. Interactive semi-automatic categorization for spinel group minerals. IEEE VAST 2015, Chicago, IL, USA, October 25-30, 2015, pp. 197–198.

[Kov18] Kovalerchuk, B. Visual knowledge discovery and machine learning, vol. 144. Springer, 2018.

[KSB+22] Kuťák, D., Selzer, M.N., Byška, J., Ganuza, M. L., Barišić, I., Kozlíková, B. & Miao, H. Vivern–A Virtual Environment for Multiscale Visualization and Modeling of DNA Nanostructures. IEEE T Vis. Comput. Gr. 2022, 28(12), 4825-4838.

[LGA+21] Luque, L. E., Ganuza, M. L., Antonini, A.S. & Castro, S.M. npGLC-Vis Library for Multidimensional Data Visualization. JCC-BD&ET 2021.

[LGC+21] Luque, L., Ganuza, M. L., Castro, S. & Agamennoni, O. E. A visualization technique to support exploratory analysis of eye movements variables in reading. En XXXV Annual Meeting of the Argentinian Society for Neuroscience Research, 2020.

[LGC+22] Luque, L., Ganuza, M. L., Castro, S. M. & Agamennoni, O. Visual analysis of eye movements during micro-stories reading. J. Vis. 2022, 25(5), 1085-1101.

[LSG+22] Larrea, M., Schiaffino, M., Ganuza, M. L. & Urribarri, D.K. A Testing Tool for Information Visualizations based on User Interactions. J. Comput. Sci. Technol. 2022, 22(1), 78-92.

[LUG23] Larrea, M.L., Urribarri, D.K. & Ganuza, M. L. New Testing Techniques to Evaluate the Quality of Information Visualization Implementations. LACCEI 2023.

[SBG+18] Splechtna, R., Beham, M., Gracanin, D., Ganuza, M. L., Bühler, K., Pandžic, I. S., and Matkovic, K. Cross-table linking and brushing: interactive visual analysis of multiple tabular data sets. Vis. Comp. 2018.

[SUS+21] Sabando, M. V., Ulbrich, P., Selzer, M., Byška, J., Mičan, J., Ponzoni, I., M. L. Ganuza & Kozlíková, B. ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening. IEEE T Vis. Comput. Gr., 2021, 27(2), 891-901.