Caracterización de Conjuntos de Datos

Dana K. Urribarri[1,2] **Silvia M. Castro**[1] **Sergio R. Martig**[1]
{dku,smc,srm}@cs.uns.edu.ar

[1] Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica (VyGLab)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Tel. 0291-4595135 Fax 0291-4595136
Bahía Blanca, CP 8000, Argentina

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Ciudad de Buenos AiresCP C1033AAJ, Argentina

RESUMEN

Contar con una taxonomía que clasifique los conjuntos de datos a visualizar es una guía que asiste a la hora de elegir la técnica de visualización apropiada para determinado conjunto de datos. Las taxonomías de datos existentes en la literatura no son presentadas desde un punto de vista de la visualización de datos, y por lo tanto ninguna de ellas proporcionan un marco que le facilite al usuario determinar qué técnica es la más apropiada. El principal objetivo de esta Línea de Investigación es la definición de una clasificación de los datos orientada a la visualización.

Palabras clave: *visualización, conjuntos de datos, taxonomías, clasificación.*

CONTEXTO

El trabajo se lleva a cabo en el Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica (VyGLab) del Departamento de Ciencias e Ingeniería de la Computación de la Universidad Nacional del Sur.

La línea de Investigación presentada está inserta en el proyecto "Interfaces No Convencionales. Su Impacto en las Interacciones" (24/Zn19), dirigido por el Lic. Sergio Martig y en el proyecto "Representaciones Visuales e Interacciones para el Análisis Visual de Grandes Conjuntos de Datos" (24/N020), dirigido por la Doctora Silvia Castro. Ambos proyectos son financiados por la Secretaría General de Ciencia y Tecnología de la Universidad Nacional del Sur; y acreditados por la Universidad Nacional del Sur, Bahía Blanca.

1. INTRODUCCIÓN

Dado el constante crecimiento de los conjuntos de datos en diferentes y variados campos de la información, la tarea de elegir la técnica más adecuada para visualizar convenientemente esos datos, no es sencilla.
Además, las técnicas y las herramientas usadas en el análisis y las visualizaciones de conjuntos de datos pequeños y medianos son inadecuadas y, en algunos casos, simplemente no son aplicables a estos grandes conjuntos de datos.

Una clasificación de los conjuntos de datos, que brinde una primera aproximación en la elección de la técnica, es de gran ayuda al momento de llevar a cabo esta tarea.
En la bibliografía existen diversas clasificaciones de los datos desde distintos puntos de vista. En [8] y [9] se presenta una clasificación básica desde un punto de vista estadístico; en [5], los datos se clasifican con el objeto de definir tareas por cada tipo de datos; en [10] se buscó generar automáticamente presentaciones visuales interactivas y presentan una taxonomía que captura las propiedades de información heterogénea, incluyendo tanto información cuantitativa como cualitativa, en entornos estáticos o dinámicos; y en [7] principalmente se desarrolla una taxonomía de visualización de alto nivel basándose en las características de los modelos de datos.
Sin embargo, ninguna de estas clasificaciones proporcionan un marco en el cual se pueda, dado un conjunto de datos, determinar qué técnica es más apropiada para su visualización, teniendo en cuenta la escalabilidad visual ([2]) de la misma y el volumen de los datos a visualizar.

2. LINEAS DE INVESTIGACION y DESARROLLO

La investigación que llevamos a cabo se centra en el desarrollo de una *clasificación* de los conjuntos de datos *orientada a la visualización*. Esta clasificación debe brindar suficiente información sobre cuales son las características con las que debe contar la técnica de visualización que se emplee para visualizar cada categoría de datos.
El desafío es encontrar métricas que, no sólo permitan evaluar en forma lo más sencilla posible cada uno de los aspectos importantes a tener cuenta, sino que también permitan una clasificación conveniente de los datos.

3. RESULTADOS OBTENIDOS/ESPERADOS

Hasta el momento hemos evaluado cuáles son las posibles métricas que se emplearán para categorizar los datos. En este sentido, se seleccionaron varias métricas globales sobre grafos, métricas sobre tablas de información y diferentes medidas de dispersión. Se buscará determinar si estas métricas, que medirán el tamaño de los conjuntos de datos, son suficientes para medir la escalabilidad visual o si además se deben incorporar características propias de la visualización en la categorización realizada.

4. FORMACION DE RECURSOS HUMANOS

En lo concerniente a la formación de recursos humanos se detallan las tesis en desarrollo y los cursos relacionados con la línea de investigación presentada dictados por los integrantes del grupo de investigación:

4.1 TESIS EN DESARROLLO

4.1.1 TESIS DE DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

- Sergio Martig. Tema: *Interacción en Visualización de Información.* Dirección: Dra. Silvia Castro.
- Martín Larrea. Tema: *Visualización basada en Semántica*. Dirección: Dra. Silvia Castro.
- Sebastián Escarza. Tema: *Ontologías de Visualización.* Dirección: Dra. Silvia Castro.
- Dana Urribarri. Tema: *Escalabilidad Visual.* Dirección: Dra. Silvia Castro.

4.2 CURSOS DE PRE Y POSGRADO RELACIONADOS CON EL TEMA DE LA LÍNEA DE INVESTIGACIÓN DICTADOS POR INTEGRANTES DEL GRUPO DE TRABAJO.

4.2.1 CURSOS DE PREGRADO

- **Computación Gráfica.** Materia optativa para los estudiantes de la Licenciatura en Ciencias de la Computación y obligatoria para los de Ingeniería en Sistemas de Computación. Universidad Nacional del Sur.
- **Comunicación Hombre-Máquina.** Materia obligatoria para los alumnos del Profesorado en Computación. Universidad Nacional del Sur.
- **Interfaces Gráficas.** Materia optativa para los estudiantes de la Licenciatura en Ciencias de la Computación y de la Ingeniería en Sistemas de Computación. Universidad Nacional del Sur.

4.2.2 CURSOS DE POSGRADO

- **Sistemas de Modelamiento de Volúmenes.** Materia del Posgrado en Ciencias de la Computación. UNS.
- **Tópicos avanzados en Curvas y Superficies.** Materia del Posgrado en Ciencias de la Computación. UNS.
- **Computación Gráfica: Tópicos Avanzados.** Departamento de Informática y Estadística de la Facultad de Economía y Administración. Universidad Nacional del Comahue.
- **Modelamiento Multirresolución.** Departamento de Informática y Estadística de la Facultad de Economía y Administración. Universidad Nacional del Comahue.
- **Visualización.** Materia del Posgrado en Ciencias de la Computación. UNS.
- **Visualización Científica.** Materia del Posgrado en Ciencias de la Computación y del Magíster en Computación Científica. UNS.
- **Visualización de Información.** Materia del Posgrado en Ciencias de la Computación. UNS.
- **Tópicos Avanzados en Visualización de Información.** Materia del Posgrado en Ciencias de la Computación. UNS.
- **Interacción Humano-Computadora.** Materia del Posgrado en Ciencias de la Computación y del Magíster en Computación Científica. UNS.
- **Modelado Geométrico Multirresolución de Superficies.** Materia del Posgrado en Ciencias de la Computación. UNS y UNLP.

5. BIBLIOGRAFÍA

[1] Currell, G., and Dowman, A. Essential Mathematics and Statistics for Science, 2 ed. John Wiley & Sons., 2009.

[2] Eick, S. G., and Karr, A. F. Visual scalability. Tech. rep., National Institute of Statistical Sciencies, 2000.

[3] Mirkin, B. Clustering for Data Mining. A Data Recovery Approach. Chapman & Hall/CRC, 2005.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1st edition, May 1999.

[5] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages (Washington, DC, USA, 1996), IEEE Computer Society, p. 336.

[6] Stuar K. Card, Jock D. Mackinglay, and Ben Shneiderman. Readings in Information Visualization Using Vision to Think. Morgan Kaufmann, 1999.

[7] Tory, M., and Moller, T. Rethinking visualization: A high-level taxonomy. IEEE Computer Society, pp. 151–158.

[8] Unwin, A., Theus, M., and Hofmann, H. Graphics of Large Datasets: Visualizing a Million (Statistics and Computing). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[9] Wegman, E. Huge data sets and the frontiers of computational feasibility. Journal of Computational and Graphical Statistics, 4 (1995), 281–295.

[10] Zhou, M. X., and Feiner, S. K. Data characterization for automatically visualizing heterogeneous information. In INFOVIS '96: Proceedings of the 1996 IEEE Symposium on Information Visualization (INFOVIS '96) (Washington, DC, USA, 1996), IEEE Computer Society, p. 13.