Caracterización de Conjuntos de Datos en Visualización*

Dana K. Urribarri
dku@cs.uns.edu.ar
Silvia M. Castro Sergio R. Martig
smc@cs.uns.edu.ar srm@cs.uns.edu.ar

Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica
Departamento de Ciencias e Ingeniería de la Computación
Tel: (0291) 459-5135 Fax: (0291) 459-5136
Universidad Nacional del Sur
Bahía Blanca, Buenos Aires, Argentina

Resumen

Contar con una taxonomía que clasifique los conjuntos de datos 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 son presentadas desde un punto de vista estadístico. La importancia de definir una clasificación de los datos orientada a la visualización radica en la necesidad de asistir a los usuarios en la elección de técnicas o estrategias de visualización que se adecúen a sus propósitos.

Palabras Claves: visualización, conjuntos de datos, taxonomía, clasificación.

1. Introducción y trabajo previo

Dada la diversidad y el voluminoso tamaño de los conjuntos actuales de datos, la tarea de elegir una técnica de visualización adecuada no es sencilla. Un método de clasificación para los diversos conjuntos de datos brinda una primer aproximación en el camino de la elección de la técnica a utilizar.

Tamaño	Descripción	Bytes
Diminuto	Entra en un pizarrón	10^2
Pequeño	Entra en unas cuantas páginas	10^4
Mediano	Entra en un disquete	10^6
Grande	Entra en un disco rígido	10^8
Enorme	Necesita varios discos rígidos	10^{10}

Tabla 1: Clasificación de Huber

Desde un punto de vista estadístico, Huber ([5]) ha planteado una clasificación (Tabla 1) que divide los datos según su tamaño. Posteriormente Wegman ([6]) extendió la clasificación para contemplar

*El presente trabajo fue parcialmente financiado por PGI 24/ZN12 y PGI 24/N020, Secretaría General de Ciencia y Tecnología, Universidad Nacional del Sur, Bahía Blanca, Argentina.

conjuntos de datos aún más grandes (Tabla 2). Sin embargo, estas dos clasificaciones tienen en cuenta solamente el tamaño en bytes del conjunto de datos y, en general, las técnicas de visualización son computacionalmente aplicables sólo dentro de los conjuntos de datos más pequeños. Una clasificación de los datos basada únicamente en su tamaño, no brinda suficiente información para elegir una técnica o estrategia de visualización acorde a los datos.

Tamaño	Descripción	Bytes
⋮	⋮	⋮
Monstruoso	Cintas magnéticas	10^{12}

Tabla 2: Extensión de Wegman

Por lo dicho anteriormente, es que resulta de suma importancia definir una clasificación de los datos *orientada a la visualización*. Una clasificación tal debe considerar aspectos adicionales más allá del tamaño en bytes, como por ejemplo cantidad de ítems de datos del conjunto, relaciones existentes entre diferentes ítems o cantidad de atributos, para que de esta forma, cada categoría brinde la información necesaria para estar en condiciones de preferir una técnica de visualización sobre otra.

2. Bases de la clasificación propuesta

Una clasificación de datos orientada a la visualización debe tener en cuenta al menos cinco aspectos: cantidad de objetos distintos, cantidad de ítems de datos, cantidad de atributos, cantidad de relaciones, y complejidad de los datos. Cada uno de estos aspectos determina alguna de las características deseables con la que debería contar la técnica más apropiada a utilizar en la visualización de dichos datos.

El desafío es encontrar métricas que, no solo permitan evaluar en forma sencilla cada uno de dichos aspectos, sino que también permitan una conveniente clasificación de los datos. Una clasificación debe ser conveniente en el sentido de que cada categoría brinde información relevante y suficiente para la visualización de los conjuntos ahí contenidos.

- Una medida que refleje la cantidad de ítems de datos permitirá determinar cuán importante es la escalabilidad de la técnica a utilizar.
- Una medida que refleje la cantidad de atributos dará la pauta de cuán necesaria es una técnica de visualización de datos multidimensionales.
- Una medida que refleje la cantidad de relaciones dirá cuán necesaria es, por ejemplo, una representación mediante grafos o si con alguna técnica que más simple es suficiente.
- La cantidad de objetos distintos a representar y la complejidad de los datos determinarán las interacciones que serán necesarias para explorar y analizar los datos.

Una taxonomía que permita diferenciar al menos estas características mínimas en los conjuntos de datos podrá dar pautas iniciales para será una gran contribución al momento de

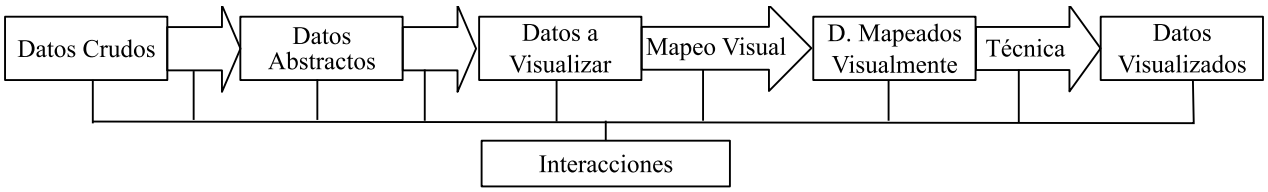

Figura 1: Modelo Unificado de Visualización

3. Clasificación en el contexto del MUV

El Modelo Unificado de Visualización (MUV [3]) refleja tanto los estados como las transformaciones intermedias que deben atravesar los datos desde que ingresan al sistema de visualización hasta que son finalmente visualizados (Figura 1).

El conjunto inicial de datos es el estado de Datos Crudos que, una vez seleccionado qué se quiere visualizar los datos pasan al estado de Datos Abstractos. En este estado se encuentran los datos potencialmente visualizables. Un subconjunto de este último conjuntos representa el estado de Datos a Visualizar, que son los datos que efectivamente estarán en la visualización. Una vez determinado el conjunto de Datos a Visualizar, la transformación de Mapeo Visual determina cómo se van a visualizar los datos: sustrato espacial, elementos visuales y atributos gráficos se emplearán en la visualización, dando lugar al estado de Datos Mapeados visualmente. Como última transformación, se aplica la Técnica (o transformación de Visualización) donde se definen demás elementos extras a la visualización de los datos (luces, colores, etc.) concluyendo en el estado final de los datos, Datos Visualizados o Vista.

Esta clasificación de los datos ayudará en el proceso de aplicación de las transformaciones de Mapeo Visual y eventualmente la transformación de Visualización. Dado un conjunto de datos, una vez determinado en qué categoría de la clasificación encuadra más acertadamente, es posible determinar con mayor facilidad qué elementos y atributos visuales son apropiados aplicar en cada caso.

4. Objetivos de la investigación

El objetivo de esta investigación es definir una clasificación de los conjuntos de datos orientada a la visualización. Una clasificación de estas características permitirá determinar técnicas de visualización aptas para cada categoría de datos, agilizando el proceso de elección de la técnica adecuada. Para cada categoría se buscarán conjuntos de datos representativos de las mismas, de forma que sea posible evaluar la validez y la efectividad de la taxonomía como soporte en el proceso de visualización.

Referencias

[1] *Modern Information Retrieval*. Addison Wesley, first edition, 1999.

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

[3] Sergio Martig, Silvia Castro, Pablo Fillottrani, and Elsa Estevez. Un modelo unificado de visualización. In *IX Congreso Argentino de Ciencias de la Computación*, octubre 2003.

[4] Jürgen Symanzik. Interactive and dynamic graphics. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics. Concepts and Methods*, pages 293–336. Springer Verlag, 2004.

[5] Antony Unwin, Martin Theus, and Heike Hofmann. *Graphics of Large Datasets: Visualizing a Million (Statistics and Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

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