Visualización de Datos Multivariados

Lidia Marina López
Dpto Informática y Estadística
Universidad Nacional del Comahue
llopez@uncoma.edu.ar
Tel 0299-4490312 Int 435 - Fax 0299-4490313

La visualización de datos multivariados implica la representación de los datos en un entorno de dimensiones mucho menores tal como una pantalla de computadora u otro medio de impresión que permita al experto inferir conclusiones. Cuando la base de datos es de un tamaño considerable, poder contar con una representación visual habilita para descubrir relaciones implícitas.

Los datos que se pretenderán visualizar corresponden al proyecto de investigación de la Facultad de Humanidades de la Universidad Nacional del Comahue: *Situación socio-profesional de género en la planta docente y órganos de gobierno de la Universidad Nacional del Comahue* y están formados por los datos personales y profesionales de dichos miembros de la Universidad.

Las técnicas de minería de datos permiten predecir situaciones de acuerdo a la distribución de la información, pero si se pudiera obtener una estructura visual, los miembros del grupo de investigación mencionado, contarían con una nueva representación que, según sus opiniones, darían un valor agregado importante.

En este marco, una línea del proyecto se está dedicando a probar las distintas técnicas de transformación de datos crudos a estructuras visuales generando la visualización propiamente dicha.

1 Presentación

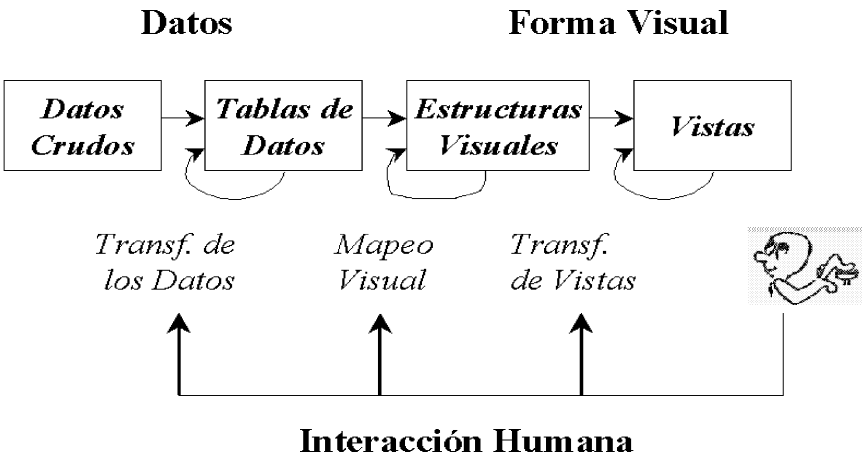
Las bases de datos personales manejan principalmente información descriptiva que, parte de ella, debe responder a un nivel de seguridad que no permita descubrir su individuo. Se debe buscar clasificarla, calcular estadísticos, cuantificar información descriptiva.

Estamos trabajando en encontrar una manera de preprocesar estos datos para luego utilizar técnicas de visualización de forma efectiva[S.C99].

Los datos crudos son transformados en relaciones o conjuntos de relaciones estructurados y fáciles de manipular para poder mapearse a formas visuales. Como este tratamiento matemático omite la información descriptiva importante para visualizar, las visualizaciones pueden pensarse como mapas ajustables que van desde datos a formas visuales, y hacia el perceptor humano. Se obtiene así, la estructura visual que representa la situación para visualizarla adecuadamente.

Mostraremos en la primera parte de este trabajo, el método de transformación de datos en *Tablas de Datos* para obtener un prototipo que mapee alguna Formas Visuales, y en la segunda parte, un breve recorrido de las técnicas de visualización multivariada. Actualmente se está trabajando sobre ambas partes y pretendemos que luego resulte en encontrar un prototipo de visualización adecuado.

2 Mapeando Datos



Este esquema muestra las transformaciones que sufren los datos crudos, ajustándose con la interacción del hombre que restringe la vista de ciertos rangos de datos o cambios en la naturaleza de la transformación. El modelo de referencia utilizado [S.C99] es el mapeado de Tabla de Datos a Estructura Visual. Las Tablas de Datos están basadas en relaciones matemáticas; las Estructuras Visuales están basadas en propiedades gráficas efectivamente procesadas por la visión humana. Si bien el dato crudo puede ser visualizado directamente, las Tablas de Datos son un importante paso intermedio para producir un componente espacial directo.

Tablas de Datos: Matemáticamente, una relación es un conjunto de *tuplas*: $\langle Value_{ix}, Value_{iy}, \ldots \rangle, \langle Value_{jx}, Value_{jy}, \ldots \rangle, \ldots$. Como este tratamiento matemático omite la información descriptiva que es importante para visualizar, se crea la noción de *Tabla de Datos*:

	$Case_i$	$Case_j$	$Case_k$	...
$Variable_x$	$Value_{ix}$	$Value_{jx}$	$Value_{kx}$	...
$Variable_y$	$Value_{iy}$	$Value_{jy}$	$Value_{ky}$	...
...	...	...	...	...

Una Tabla de Datos combina relaciones con *metadatos* que describen aquellas relaciones. Las etiquetas para las filas y las columnas son un ejemplo de metadatos. Las filas representan *variables*, para los rangos de los valores en las tuplas. Las columnas representan los casos, que son conjuntos de valores para cada una de las variables. La doble línea vertical distingue una Tabla de Datos de otros tipos de tablas. El orden de las filas y las columnas en una Tabla de Datos puede o no ser significativo. Este orden es otro ejemplo de metadatos que es importante para visualización.

Una de las ventajas de las Tablas de Datos es que claramente describen el número de variables asociado con un conjunto de datos, una consideración importante cuando se seleccionan visualizaciones. La dimensionalidad se usa para referirse al número de variables de entrada, al número de variables de salida, a ambos, o más aún, al número de dimensiones espaciales en los datos.

Transformaciones de Datos: A menudo los datos crudos contienen errores o valores pérdidos que deben ser direccionados antes de que los datos puedan ser visualizados. Los cálculos estadísticos pueden también agregar información adicional. Por estas razones, las Tablas de Datos contienen valores o estructuras derivadas. Existen cuatro tipos de estas transformaciones de datos:
$Valor \rightarrow Valor\ Derivado$
$Estructura \rightarrow Estructura\ Derivada$
$Valor \rightarrow Estructura\ Derivada$
$Estructura \rightarrow Valor\ Derivado$

Las transformaciones de datos pueden ser concatenadas para formar cadenas de agregación y clasificación como parte del proceso de cristalización del conocimiento. Se pueden descubrir patrones y agregarlos como un nuevo dato del esquema codificándolos en las variables de la Tabla de Datos. Las visualizaciones de Tabla de Datos pueden ser usadas para detectar más patrones. Los controles operados por el usuario sobre transformaciones estructurales de la Tabla de Datos pueden ser usados como controles sobre la visualización.

3 Visualizando Datos

Muchas técnicas para representar Estructuras Visuales han sido desarrolladas a partir de 1987 [Nie97]. Estas técnicas combinan las métricas espaciales, las marcas y las propiedades gráficas de manera realmente exitosa consiguiendo resultados excelentes. De acuerdo a estas combinaciones, las hemos clasificado en: *Representaciones Matriciales* (composición de ejes ortogonales)
Iconografía (métricas espaciales y propiedades gráficas)
Representaciones Jerárquicas (conexión y contenido)
Otras Representaciones (Brushing y representaciones no cartesianas)

Representaciones Matriciales Se refieren a visualizaciones que codifican información posicionando marcas sobre ejes ortogonales. El desarrollo más importante aquí es el análisis de datos exploratorio de Tukey [Tuk77], pilar importante en la visualización de datos. Raramente se utiliza color. Muchas de estas herramientas muestran correlaciones entre dos variables. Aportes importantes se realizaron durente los 80 con Cleveland [S.C93].

Iconografía Esta técnica utiliza pequeños objetos gráficos, llamados íconos o glyfos, Este término se utiliza para describir cualquier entidad gráfica que mapea puntos de datos. El glyfo tiene atributos varios tales como tamaño, forma, y color controlados por el valor del punto de información en dimensiones diferentes.[War94][R.P98] Los glyfos proveen una manera efectiva de visualizar arriba de 9 atributos concurrentemente: 3D posición, 3D tamaño, color, forma, opacidad.

Representaciones Jerárquicas El *apilado dimensional* es una técnica donde las dimensiones están dentro de otras dimensiones. El orden más alto de dimensión divide la representación en regiones rectangulares. Dentro de cada una de estas regiones, la siguiente dimensión más alta divide la representación en más regiones rectangulares, y así sucesivamente.[JN90] Otra técnica importante es *Mundos dentro de Mundos*, método que explora e intenta entender la visualización de funciones de muchas variables. [SC90]

Para grafos existen algoritmos que producen buenas representaciones tanto top-down como left-right de grafos en forma arbórea [EJ81][II90]. Una técnica popular es representar grafos en 3D en lugar de 2D con la esperanza de que la dimensión extra brinde, literalmente, más espacio y puedan representarse grandes estructuras. La representación hiperbólica (principalmente árboles) es una de las formas nuevas de representar grafos que ha sido desarrollada teniendo en cuenta visualización e interacción. Los primeros trabajos se desarrollaron sobre visualizadores del contenido de la Web.[JP95][JR96]

Otras Representaciones -*Brushing*, definición estándar aceptada para *brush*: entidad gráfica que contiene un subconjunto de datos que está siendo visualizado y controlado por el usuario en una manera rápida, intuitiva e interactiva.[BC87][MW95]

-*Coordenadas Paralelas*, en este método cada dimensión corresponde a uno de los ejes verticales espaciados uniformemente. Un punto de dato mapea a un conjunto de puntos a lo largo de cada eje. El punto se representa dibujando una línea de todos los ejes conectando los puntos. [AB90][War94]

4 Estado actual y trabajo futuro

El proyecto mencionado anteriormente, donde pretendemos aplicar la visualización, tiene como objetivos específicos mostrar la distribución por género de cargos y funciones de la planta docente y órganos de gobierno, establecer modelos de relaciones existentes entre las diferentes jerarquías según las normas estatutarias y las prácticas establecidas para poder formar el entramado de relaciones socio-educativas al interior de la Universidad.

El estado actual de la rama de visualización se encuentra en la etapa de construcción de la Tabla de Datos y análisis y evaluación de las distintas técnicas de representación. En estos momentos hemos conseguido lograr representaciones de hasta seis dimensiones con algunas de las técnicas mencionadas. El objetivo final es lograr una visualización de por lo menos nueve dimensiones.

Referencias

[AB90] A.Inselberg and B.Dimsdale. *Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry, IEEE Visualization 90.* IEEE Computer Society Press, Los Alamitos, California, 1990.

[BC87] Richard Becker and Williams Cleveland. Brushing scatterplots. *Technometrics*, 29:127–142, 1987.

[EJ81] E.M.Reingold and J.S.Tilford. Tidier drawing of trees. *IEEE Trans. Software Engeniering*, 7(2):223–228, 1981.

[II90] Q.Walker II. A node-positioning algorithm for general trees. *Software -Practice and Experience*, 20(7):685–705, 1990.

[JN90] M.Ward J.LeBlanc and N.Wittels. Exploring n-dimensional databases, 1990.

[JP95] R.Rao J.Lamping and P.Pirolli. A focus+context technique based on hyperbolic geometry for visualizating large hierarchies. In *Human Factors in Computing Systems*. CHI 95 Conf.Proc., 1995.

[JR96] J.Lamping and R.Rao. The hyperbolic browser: A focus+context technique for visualizating large hierarchies. pages 33–55. J. Visual Languages and Computing - vol7, no 1, 1996.

[MW95] Allen R. Martin and MAtthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceeding of IEEE Visualization 95*, pages 271–278, Los Alamitos, California. Oct 1995, 1995. Gregory M. Nielson and Deborah Silver, IEEE Computer Society Press.

[Nie97] Muller Nielson, Hagen. *Scientific Visualization. Overviews-Methodologies-Techniques*. IEEE Computer Society, California, USA, 1997.

[R.P98] G.Grinstein R.Pickett. Iconographics displays for visualizating multidimensional data, 1998.

[SC90] S.Feiner and C.Beshers. Visualizing n-dimensional virtual worlds with n-vision, 1990.

[S.C93] William S.Cleveland. *Visualization Data*. Hobart Press, Summit, NJ, 1993.

[S.C99] B.Shneiderman S.Card, J.Mackinlay. *Readings in Information Visualization - Using Vision to Think*. Morgan Kaufmann Publisher, Inc, San Francisco, California, 1999.

[Tuk77] John W. Tukey. *Exploratory Data Anaysis*. Addison-Wesley, 1977.

[War94] Matthew O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceeding of IEEE Visualization 94*, pages 326–336, Los Alamitos, California. Oct 1994, 1994. Daniel Bergeron and Arie E. Kaufman, IEEE Computer Society Press.