Escalabilidad visual en coordenadas paralelas

Dana K. Urribarri
Silvia M. Castro
Sergio R. Martig
Laboratorio de Visualización y Computación Gráfica
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Bahía Blanca, CP 8000, Argentina

RESUMEN

En sistemas de áreas tan diversas como simulación, salud, la Web, meteorología, o testeo de productos, los volúmenes de datos son cada vez mayores y se agigantan constantemente. A la hora de analizar e interpretar estos datos las limitaciones humanas y la falta de software adecuado para complementarlo, son los mayores problemas. Esta carencia de software se debe principalmente a la complejidad computacional que implica procesar tales conjuntos de datos y además a que las técnicas de visualización eficaces para volúmenes de datos reducidos no son aplicables en estos casos. El desarrollo de técnicas escalables visualmente es sustancial a la hora de producir herramientas adaptables a conjuntos de información de gran magnitud. El principal objetivo del trabajo de investigación que se está desarrollando es realizar el análisis de la factibilidad de escalar visualmente las coordenadas paralelas, que es una de las técnicas más poderosas de visualización *n*–dimensional. De este modo se podrá extender su uso a visualizaciones de grandes conjuntos de datos.

Palabras claves: visualización de información, escalabilidad visual, coordenadas paralelas.

1. INTRODUCCIÓN

El tamaño de los volúmenes de datos provenientes de distintas aplicaciones tanto científicas como no científicas crece constantemente y a pasos agigantados. Es por esto que resulta imprescindible contar con herramientas de visualización de información que proporcionen las facilidades necesarias para que el usuario pueda interpretar correctamente los datos, y por lo tanto, tomar decisiones adecuadas basándose en ese análisis. En general, las técnicas de visualización para conjuntos de datos pequeños o medianos no escalan correctamente a grandes conjuntos de datos, es por esto que diversas interacciones pueden contribuir a que la técnica sea aplicable a estos grandes conjuntos.

Muchos trabajos, como consecuencia de la complejidad computacional, se han enfocado en la escalabilidad de los algoritmos. Sin embargo, la escalabilidad de la visualización en sí no es menos importante [1]. La escalabilidad visual es la capacidad de las herramientas de visualización de mostrar efectivamente grandes conjuntos de datos, ya sea por la cantidad de ítems o por la dimensionalidad de éstos. Sin embargo, hay implementaciones de técnicas tales como gráficos de barras, scatterplots o grafos que al aplicarse a conjuntos de datos grandes, no escalan efectivamente.

2. UNA TÉCNICA EN PARTICULAR

Una técnica de gran importancia para visualización de conjuntos de datos multidimensionales es la de las coordenadas paralelas [2]. Informalmente, esta técnica consiste en asignarle a cada dimensión un eje y disponer estos ejes paralelamente en el plano. Cada dato *n*-dimensional (a_1, a_2, a_3,…, a_n) es una poligonal que atraviesa los *n* ejes paralelos en los puntos (p_1, p_2, p_3,…, p_n) (Figura 1).

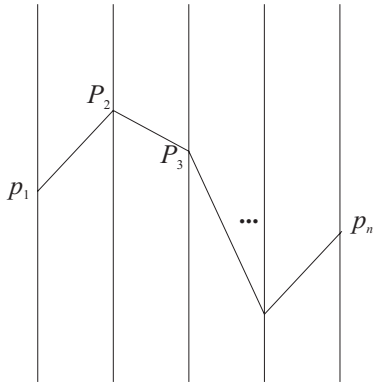

Figura 1

Además de ser una técnica apta para datos multidimensionales, es también apropiada para grandes conjuntos de datos.

Problemas que solucionan

Las coordenadas paralelas hacen posible visualizar datos de tantas dimensiones como sea necesario: la clave está en no representar las dimensiones con ejes ortogonales, sino con ejes paralelos. Esto no sólo logra una representación uniforme para todas las dimensiones, ya que todas las dimensiones se

representan con ejes, sino también logra que agregar una dimensión más a los datos no provoque un cambio radical en la forma de representación. Además, la detección de patrones de comportamiento en los datos dentro de un espacio *n*-dimensional, se reduce a interpretar un gráfico en dos dimensiones.

Problemas inherentes a la técnica

A pesar de sus grandes ventajas, las coordenadas paralelas también presentan algunos inconvenientes, entre los que se encuentran principalmente la disyuntiva *muchos atributos – espacio acotado* y la *oclusión*.

El primer problema es consecuencia de que el número de dimensiones que se pueden representar es potencialmente ilimitado, aunque el espacio que se dispone para la representación (la pantalla del monitor) es acotado y de resolución finita. Esto implica que:

1- la cantidad de ejes que se pueden graficar está acotada por la resolución (vertical u horizontal, dependiendo de cómo se dispongan los ejes),
2- los ejes deben estar lo suficientemente espaciados para que sea posible distinguir cuál es el comportamiento de las poligonales.

El segundo gran problema es la *oclusión*. Si al problema de la resolución finita se le suma la gran cantidad de poligonales, el resultado obtenido es una gráfica incomprensible, donde es imposible individualizar los datos o peor aún, es imposible distinguir patrones que insinúen el comportamiento de estos datos.

Dado que estos problemas son inherentes a la técnica y crecen con la cantidad de datos a visualizar y, por otro lado esta técnica es de gran utilidad, se hace necesario proveer al usuario con herramientas que le permitan solucionarlos. De este modo se permitirá lograr la escalabilidad visual adecuada para visualizar grandes conjuntos de datos. Una de las posibles soluciones es, sin duda, brindar al usuario interacciones adecuadas, así este podrá explotar al máximo el potencial de la técnica frente a grandes volúmenes de datos.

3. TRABAJO PREVIO

Varios autores han aportado diferentes modificaciones o interacciones que intentan ayudar al usuario en su tarea de analizar los datos, ya sea intentando solucionar alguno de los problemas propios de las coordenadas paralelas o diseñando facilidades extras.

Se han presentado diferentes alternativas para sobreponerse a la oclusión. Se han presentado coordenadas paralelas jerárquicas [3] como una estrategia de clusterización. Es una variación del gráfico de coordenadas paralelas que ofrece una vista multirresolución de los datos a través de clustering jerárquico (Figuras 2 y 3).

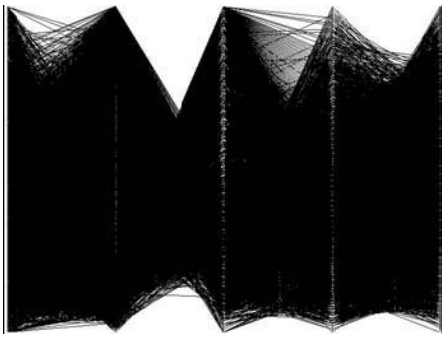

Figura 2: Un conjunto de datos visualizado con Coordenadas Paralelas

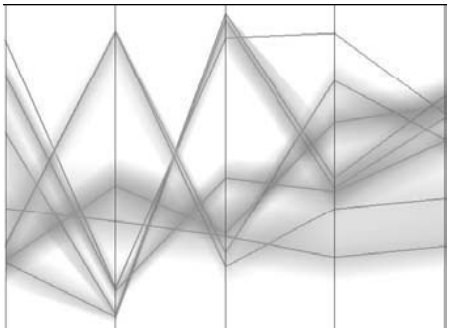

Figura 3: El mismo conjunto de datos visualizado con Coordenadas Jerárquicas

Otra estrategia planteada ha sido generar gráfico de frecuencias [4], que se basa en calcular para cada par de atributos que se mostrarán en forma adyacente la frecuencia de cada segmento de dato, y colorear el segmento en función de este valor. La mayor desventaja está en que las poligonales no se colorearán en forma uniforme, haciendo que se pierda la continuidad de los datos a lo largo de la gráfica. Para evitar esto presentan algunas interacciones necesarias.

Otro problema, aunque no tan grave como la oclusión cuando se trata de grande volúmenes de datos, es el entrecruzamiento de poligonales [5]. Cuando dos o más poligonales se encuentran en el mismo punto sobre un eje, para los demás ejes, es imposible distinguirlas entre sí (Figura 4a). La solución planteada no representa los datos como poligonales, sino como curvas suaves (Figura 4b).

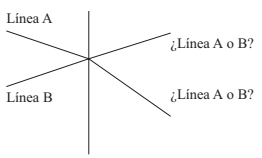

Figura 4a

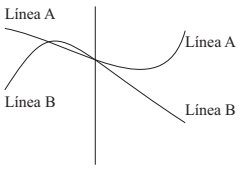

Figura 4b

También se ha trabajado sobre interacciones adicionales a la técnica. En [6] se presentan el brush angular, brush suave y composición de brushes. El *brush angular* permite visualizar correlaciones

positivas o negativas entre los atributos, basándose en la pendiente de los segmentos entre dos ejes adyacentes. El *brush suave* asigna al brush una función de grado de interés (DOI) entre 0 y 1, y cuanto más alejada una poligonal del centro de interés, más cercano a 0 es el DOI. La *composición de brushes* implica definir varios brushes independientes y operaciones lógicas (and, or, not) entre brushes (Figura 5).

En [7] intentaron encontrar el escalado de los ejes apropiado dependiendo del propósito de la visualización. Se distinguen diferentes tipos de tareas: investigar características de los objetos y relación entre atributos, análisis de similaridad entre datos o análisis de costo/beneficio. Para cada una de estas tareas plantean un escalado y alineación diferente de los ejes.

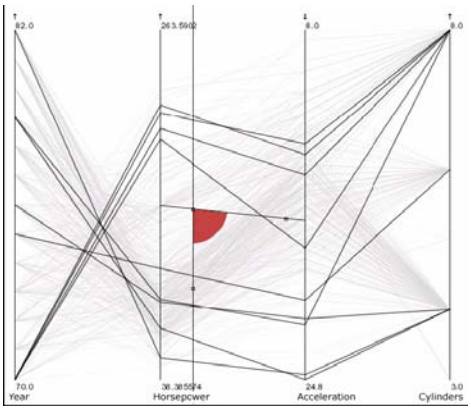

Figura 5

4. OBJETIVO DE LA INVESTIGACIÓN

El objetivo del trabajo a realizar es analizar y definir las diferentes interacciones que son necesarias sobre un gráfico de coordenadas paralelas para hacer de ésta una técnica escalable y así aprovechar al máximo su potencial para visualización de grandes volúmenes de datos. La idea es determinar la escalabilidad visual de la técnica en función de las interacciones que se pueden realizar sobre ella. Por otro lado, dada la importancia de las interacciones en el proceso de visualización se pretende clasificarlas de acuerdo a su funcionalidad y al impacto que éstas tienen sobre el Modelo Unificado de Visualización [8].

El Modelo Unificado de Visualización refleja los estados de los datos desde que ingresan al sistema de visualización hasta que son finalmente visualizados, y las transformaciones intermedias que hacen posible la evolución de los datos a lo largo de los diferentes estados. Tomando este modelo como referencia, se encuadrarán las interacciones en el mismo; esto nos provee un marco teórico común como base para analizar y comparar diferentes técnicas de visualización. En nuestro caso particular, es una manera de comparar, en cuanto a la escalabilidad visual, las coordenadas paralelas con otras técnicas de visualización apropiadas para grandes conjuntos de datos.

5. AGRADECIMIENTOS

El presente trabajo fue parcialmente financiado por PGI 24/N015, Secretaría General de Ciencia y Tecnología, Universidad Nacional del Sur, Bahía Blanca, Argentina.

6. REFERENCIAS

[1] Stephen G. Eick and Alan F. Karr. Visual Scalability. Technical Report Number 106. June, 2000.

[2] Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. *IEEE Visualization*, pages 361–378, 1990.

[3] Ying-Huey Fua, Matthew O.Ward, and Elke A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. *IEEE Visualization*, pages 43–50.

[4] Almir Olivette Artero, María Cristina Ferreira de Oliveira, and Haim Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. *IEEE Symposium on Information Visualization 2004*, pages 81–88, October 2004.

[5] Martin Graham and Jessie Kennedy. Using curves to enhance parallel coordinate visualisations.

[6] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular brushing of extended parallel coordinates.

[7] Gennady Andrienko and Natalia Andrienko. Constructing parallel coordinates plot for problem solving. *In Proc. 1st International Symposium on Smart Graphics*, pages 9–14, 2001.

[8] Sergio Martig, Silvia Castro, Pablo Fillottrani, and Elsa Estévez. Un modelo unificado de visualización.