Visualización de datos multidimensionales en la Web guiada por los datos y por el usuario*

Rober Morales-Chaparro, Andrés Iglesias-Pérez, Juan Carlos Preciado

Quercus Software Engineering Group, Universidad de Extremadura
{robermorales,andresip,jcpreciado}@unex.es

Resumen La visualización de datos en la Web es una de las herramientas más populares para interpretar la información proveniente de los sistemas de *Business Intelligence*. Sin embargo, la variedad de fuentes de datos y dispositivos junto con la naturaleza multidimensional de los datos y la evolución continua de las necesidades de la visualización están haciendo a esta disciplina cada vez más desafiante. En este trabajo se describe un procedimiento para obtener una visualización de datos multidimensionales conducida tanto por los datos como por el usuario, proporcionando una generación automática de código.

Keywords: Visualización de datos,Ingeniería Web,Interacción Hombre-Máquina

1. Introducción

La visualización de datos está siendo cada vez más utilizada en las aplicaciones de *Business Intelligence* basadas en Web. No sólo por la importancia de extraer información relevante para las empresas, sino por la dificultad de mostrarla adecuadamente. Los directivos quieren ver el estado de sus negocios de manera fácil y rápida, para poder tomar decisiones correctas y a tiempo [Bro08,TSD10,Rob08].

Estas técnicas se han manifestado como útiles si los requisitos no cambian con el tiempo. Sin embargo, para aquellas aplicaciones con requisitos que evolucionan, estos sistemas se muestran limitados. La solución deseable para este problema incluye una mayor participación del usuario. Varios autores han identificado este reto: debe haber más equilibrio comunicativo entre los expertos en visualización y los que dominan el problema [KEM07]. Por otro lado, la variedad de fuentes de datos (lenguajes de consulta, APIs, tecnologías,...) junto con la gran diversidad de dispositivos para visualización y las últimas prácticas popularizadas en las aplicaciones sociales (etiquetado, valoración, sugerencias,...) están haciendo que la visualización en la Web tenga cada vez más retos que acometer.

La mayor responsabilidad de una interfaz de usuario es hacer un uso inteligente de las habilidades humanas para percibir información. El objetivo en este sentido es por tanto encontrar metáforas visuales efectivas para los datos. Por ejemplo,

* Este trabajo ha sido desarrollado bajo el proyecto TIN2008-02985 y con ayuda de Junta de Extremadura y FEDER

las metáforas útiles para representar una variable cuantitativa (como pueda ser el tamaño) no son las mismas que para una cualitativa (como el color). Además, la visualización difiere dependiendo también de la dimensión dominante.

La principal contribución de este trabajo es la presentación de un proceso de construcción de visualizaciones dirigido por los datos y por el usuario. Aporta dos beneficios. El primero, la posibilidad de que el diseñador reutilice diferentes patrones de visualización entre diferentes aplicaciones. El segundo, la posibilidad para el usuario de disponer de varias visualizaciones distintas para el mismo conjunto de datos. Este trabajo tan solo presenta las líneas generales.

El artículo está organizado de la siguiente manera: después de esta introducción explicaremos con un ejemplo la motivación del trabajo. A continuación explicaremos la propuesta desde los datos, pasando por los modelos y llegando al usuario. Al fin, revisaremos los trabajos relacionados y las conclusiones.

2. Ejemplo

La gestora de los restaurantes de un aeropuerto internacional usa un sistema de visualización que le resume los datos sobre los vuelos (Figura 1). Esta solución se adaptaba a sus necesidades originalmente. Ahora, sin embargo, está pensando en abrir restaurantes temáticos en los que la carta dependa de la nacionalidad de la mayoría de los usuarios del aeropuerto, según el horario. Por tanto, necesita saber qué países son los orígenes/destinos de los vuelos en diferentes momentos del día. Ella conoce qué información quiere y cómo la quiere visualizar.

Sin embargo, su sistema de visualización no es capaz de mostrarle los datos en la forma que desea. Necesita un sistema que conozca la naturaleza de los datos como para ofrecerle la posibilidad de elegir entre diferentes visualizaciones. Este sistema será presentado en la siguiente sección.

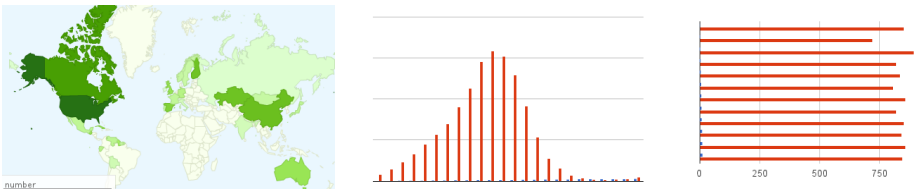

Figura 1. Monitor original. Izquierda: vuelos por país de destino/origen. Centro: vuelos por hora. Derecha: vuelos por mes.

3. Propuesta

La propuesta tiene tres partes: sistema de extracción de datos, sistema de selección de visualizaciones y sistema de generación de código (Figura 3). A continuación explicamos las tres partes.

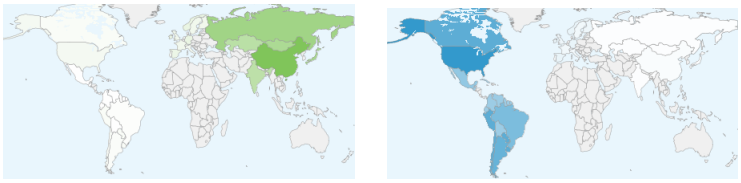

Figura 2. Monitor deseado. A las 9h y a las 15h, frecuencia de vuelos por país.

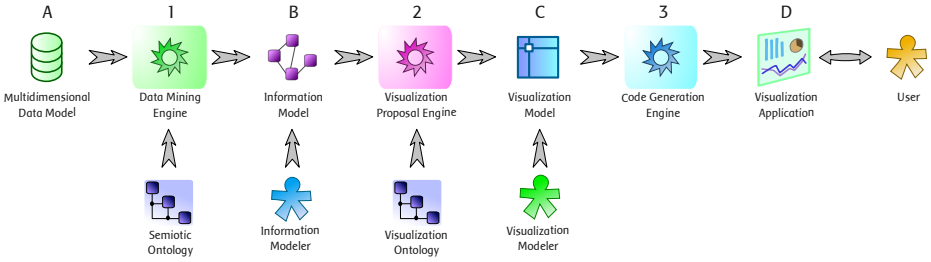

Figura 3. Pasos secuenciales de la propuesta.

3.1. Guiado por los datos

Las dimensiones son los atributos de los objetos representables del conjunto de datos (como la edad), o bien agregaciones de estos (como la edad media). Están marcadas semánticamente con (a) el tipo, (b) el rango si es aplicable, (c) la importancia, (d) la relación con otras dimensiones y (e) el nivel organizacional (básicamente: cualitativo $\subset$ ordinal $\subset$ cuantitativo).

El motor de minería de datos (Figura 3-1) es el punto de entrada del proceso. El objetivo del mismo es proponer perspectivas de los datos que maximicen (a) la utilidad y (b) la precisión de percepción. Para obtener estos dos objetivos, automatiza cada aspecto que se pueda inferir desde los datos. Toma como entrada el modelo de datos anotado (Figura 3-B) y da como salida uno de información (Figura 3-C), ayudándose de una ontología semiótica, no mostrada, que contiene las valoraciones entre la información semántica y las variables visuales.

La primera tarea de la fase es detectar las perspectivas más relevantes: esto incluye elegir la dimensión dominante, y si hay que agrupar la información. La segunda tarea incluye la búsqueda de la correspondencia entre estas perspectivas y los patrones de visualización disponibles. Esto hay que hacerlo teniendo en cuenta la dimensión dominante, pero también el resto de variables, el tamaño del conjunto de datos, etc. Para nuestro ejemplo, el modelo de datos anotado se presenta en la Figura 4.

El cuadro 1 muestra el resultado de la primera tarea: las perspectivas que el análisis de minería de datos ha encontrado relevantes. El cuadro 2 es el resultado de la siguiente tarea: las mejores metáforas visuales para las dimensiones de esas perspectivas. La fase conducida por los datos finaliza con un modelo de información no mostrado (que sería la composición de los cuadros 1 y 2).

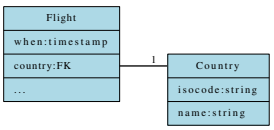

Figura 4. Modelo de datos anotados.

Cuadro 1. Perspectivas propuestas.

Perspectiva	**dimensiones**		
	dominante	**salidas**	**group**
vuelos por hora	when	*count*	hour
vuelos por mes	when	*count*	month
vuelos por país	Country	*count*	hour

Cuadro 2. Correspondencia entre atributos y variables visuales.

Dimensiones →	**Nivel organizational** →	**Variables**
count	cuantitativo	intensidad, tamaño
when	ordinal	eje horizontal
country	cualitativo (geo)	mapa

3.2. Dirigido por modelos

En este punto, el diseñador tiene la opción de modificar el modelo de información propuesto por el sistema. Después de esto, el motor de selección de visualizaciones elegirá los mejores patrones para mostrar estas perspectivas (Figura 3-2). La salida del sistema es un esqueleto de modelo de visualización (Figura 3-C) que el modelador puede refinar. Los patrones se almacenan en una ontología, para que se puedan utilizar para diferentes conjuntos de datos. El sistema de generación de código (Figura 3-3) está pensado para transformar todos los modelos en un sistema ejecutable.

En nuestro ejemplo, podemos ver en el cuadro 3 la puntuación de diversos patrones para las perspectivas. El sistema propone los patrones mejor puntuados.

El proceso propuesto, en este punto, ha detectado que el mejor monitor para el sistema es el que la gestora del aeropuerto ya tiene. Ahora vamos a ver otras ventajas adicionales.

3.3. Guiado por el usuario

Usando el almacenamiento formal que nos aportan las ontologías, así como la representación también formal de los modelos, el sistema puede, en tiempo de ejecución, permitir al usuario de lo siguiente: a) manteniendo los patrones de visualización, editar las preferencias (color, disposición, etc.); b) probar para un mismo conjunto de datos otros patrones (manteniendo la dimensión dominante) y después, opcionalmente, puede hacer a); c) cambiar la dimensión dominante y

Cuadro 3. Resultados de diferentes patrones evaluados para mostrar las perspectivas seleccionadas.

	bar chart	column chart	line chart	pie chart	heat map	time line	⋰
vuelos por hora	**0.8**	0.7	0.5	0.4	0.0	0.3	...
vuelos por mes	**0.8**	0.7	0.5	0.4	0.0	0.3	...
vuelos por país	0.5	0.4	0.0	0.3	**0.9**	0.0	...

obtener por tanto nuevas perspectivas y a continuación, b) y a); d) cambiar el conjunto de datos que quiere ver, y después opcionalmente c), b) y a).

Ahora podemos revisitar el ejemplo desde alguno de estos casos, supongamos el c). Tal y como se explicó, la gestora del aeropuerto quería ver los "vuelos por hora y país". Una vez expresado esto, el sistema de minería de datos encontrará las mejores metáforas visuales para estos datos, y el sistema de selección de visualizaciones buscará los mejores patrones para esas variables.

Siguiendo el proceso descrito en este trabajo se ha desarrollado una aplicación de demostración accesible desde `http://visualligence.com/airport_gv/`.

4. Conclusiones y trabajos relacionados

Cuadro 4. Resumen de los trabajos relacionados.

	Usuario	**Datos**	**MDE**	**Multidim**	**Herramienta**	**Web**
[MLG+10]	Sí	Sí	No	No	Sí	No
[The03]	Sí	No	No	Sí	Sí	No
[BSL+01]	Sí	No	No	Sí	Sí	No
[SCB98]	Sí	No	No	Sí	Sí	No
[FPSSO96]	No	Sí	No	Sí	Sí	No
[Mac86]	No	Sí	No	Sí	Sí	No
[SH00]	Sí	No	No	Sí	Sí	No
[BBC+11]	No	Sí	Sí	Sí	Sí	Sí
Nuestra propuesta	Sí	Sí	Sí	Sí	Sí	Sí

En este trabajo se ha presentado una propuesta guiada por los datos y por el usuario para generar visualizaciones para la Web, en base a modelos. Huyendo de los detalles, se han contado las principales ideas del proceso y sus fases. Este proceso está en fase de integración con RUX [LPSF07], una herramienta basada también en modelos para generar interfaces ricas de usuario.

Se ha realizado una evaluación de características en diferentes trabajos relacionados (¿guiado por el usuario?, ¿por los datos?, ¿basado en modelos?,

¿multidimensional?, ¿con herramienta?, ¿orientado a la Web?). El cuadro 4 muestra el resultado de la misma, y bajo nuestro conocimiento no hay ningún trabajo que considere todos estos aspectos a la vez.

Referencias

BBC+11. A Bozzon, M Brambilla, T Catarci, S Ceri, P. Fraternali, and M Matera. Visualization of Multi-domain Ranked Data. *Search Computing*, pages 53–69, 2011.

Bro08. Michael G Brooks. The Business Case for Advanced Data Visualization, 2008.

BSL+01. Andreas Buja, D.F. Swayne, M. Littman, Nathaniel Dean, and H. Hofmann. XGvis: Interactive data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, pages 1061–8600, 2001.

FPSSO96. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and Others. Knowledge discovery and data mining: Towards a unifying framework. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR*, pages 82–88, 1996.

KEM07. A. Kerren, A. Ebert, and J. Meyer. *Human-Centered Visualization Environments*. Springer-Verlag Berlin Heidelberg, 2007.

LPSF07. Marino Linaje, Juan Carlos Preciado, and Fernando Sánchez-Figueroa. Engineering Rich Internet Application User Interfaces over Legacy Web Models. *IEEE Internet Computing*, 11(6):53–59, November 2007.

Mac86. Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, April 1986.

MLG+10. K. Matkovic, A. Lez, D. Gracanin, Andreas Ammer, and Werner Purgathofer. Event Line View: Interactive Visual Analysis of Irregular Time-Dependent Data. In *Smart Graphics: 10th International Symposium on Smart Graphics, Banff, Canada, June 24-26 Proceedings*, page 208. Springer Verlag, 2010.

Rob08. J.C. Roberts. *The Craft of Information Visualization*. January 2008.

SCB98. Deborah F. Swayne, Dianne Cook, and Andreas Buja. XGobi: Interactive Dynamic Data Visualization in the X Window System. *Journal of Computational and Graphical Statistics*, 7(1):113, March 1998.

SH00. C. Stolte and P. Hanrahan. Polaris: a system for query, analysis and visualization of multi-dimensional relational databases. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 5–14, 2000.

The03. Martin Theus. Interactive data visualization using mondrian. *Journal of Statistical Software*, 7(11):1–9, 2003.

TSD10. E. Turban, R. Sharda, and D. Delen. *Decision support and business intelligence systems*. Prentice Hall Press Upper Saddle River, NJ, USA, 2010.