Frameworks para visualización y análisis de volúmenes de datos en Big Data usando modelamiento de ecuaciones estructurales

Walter Hugo Arboleda Mazo - UNAC
Leydy Johana Orozco Carvajal - Universidad Católica Luis Amigó

Resumen

En las ciencias computacionales, la visualización de información con frecuencia requiere el uso de métodos sofisticados, de forma que se pueda mostrar información importante, resultado de la relación entre las variables. Es así como mediante proyecciones y regresiones usando modelamiento de ecuaciones estructurales, se pueden tomar decisiones y ver aspectos no visibles en otro tipo de visualizaciones; para la realización de esta forma de obtener conocimiento, se usan frameworks y herramientas que reciben como insumo los valores entregados por las variables, haciendo posible para el usuario final que el análisis de información de forma dinámica e interactiva sea cada vez más fácil.

Palabras clave: modelamiento de ecuaciones estructurales, decisiones, correlación, visualización de información, análisis de información, big data.

Abstract

In Computational sciences, information visualization requires often use sophisticated methods, in order to show sensitive information, as a result of the relationship among variables. This way using structural equation modeling for projections and regressions, it is possible to take decisions and observe non-visible aspects in another kind of visualization; to create these kinds of knowledge acquiring, some type of frameworks and tools are used in order to receive data from variables, this way is possible for the final user to analyze information easily in a dynamic and interactive form.

Keywords: structural equation modeling, decisions, correlation,

information visualization, information analysis, big data.

1. Introducción

En la visualización de volúmenes de datos es importante saber qué fenómeno se desea analizar, para decidir cuál es la mejor forma de visualización, apareciendo interrogantes como los siguientes: ¿Cuál será la audiencia que hará análisis de los datos?, ¿qué interrogantes se tienen con respecto a lo que se espera resolver en el análisis y visualización?, ¿qué respuestas se buscan al analizar la información?, ¿qué inferencias pueden surgir a partir de visualizar los datos de una forma fácil y rápida? Estas preguntas llevan a seleccionar el tipo de método que ayudará a alcanzar dicho objetivo, según la naturaleza de los datos, y luego definir qué tipo de gráficos se requieren para mostrar los resultados de forma que se pueda interactuar con la gráfica interactivamente, pasando desde gráficos de líneas, barras, columnas y circulares a nuevas formas de mostrar información, como los análisis de correlación, los cuales pueden ser fácilmente visualizados usando Scatter plot.

Se facilita que con frecuencia se construyan tableros con diversas opciones de visualización y mostrando diversas correlaciones entre las variables, dando varias perspectivas de la misma información, haciendo más fácil un análisis dinámico de información que permita realizar conclusiones rápidamente, lo que de otra forma sería complejo por la cantidad de variables que se requeriría recalcular en tiempo real.

2. Uso de estadística multivariada en la visualización y análisis de volúmenes de datos

El término Estadística Multivariada (EM) se usa cuando se tienen más de dos variables simultáneamente analizadas (Wuensch, 2016); para la creación de modelos de EM se requieren una matriz de datos y una matriz de correlación (covarianzas entre las variables), lo que exige condiciones dependiendo del caso de análisis, el cual está asociado a los datos que serán analizados, cómo se relacionan estos y cómo pueden generar información.

Para hacer análisis de datos existen métodos como Factor Analysis (FA), Multiple Regression (MR), Multivariate Analysis of Variance (MANOVA) y Analysis of Covarianza (ANCOVA), los cuales se centran en la covarianza, el radio de las varianzas, permitiendo interpretar las combinaciones lineales entre los componentes y los factores, las pruebas de significancia y las medias o los pesos.

Un ejemplo de uso de MR y FA, los cuales son la base de SEM, son los análisis usados en minería de datos de datos sociales para evaluar el concepto de calidad de vida y su influencia en los asentamientos humanos, facilitando determinar la relación entre la satisfacción de los ciudadanos y las variables latentes: calidad ambiental, costos de

producción, servicios públicos básicos, costo de vida, entretenimiento, atmósfera social e imagen de la ciudad (Kan, Xuefei & Jin, 2014).

También se usan estos métodos para evaluar la selección de carrera en programa de IT, lo que depende de las variables latentes o constructos: habilidades del aspirante en IT, expectativas al terminar la carrera (Luse, 2014), variables las cuales a la vez influyen directamente en la variable llamada interés en IT.

Se permite así, mediante estudios y mediciones empíricas, llegar a construir modelos y validar teorías e hipótesis, usándose grandes volúmenes de datos que sean representativos, se puedan usar en el modelo y representen la realidad, haciendo que el modelo creado sea efectivo (Templin, 2011). Este paso permite entender que el análisis multivariado de la varianza posibilita analizar múltiples respuestas continuas; la regresión logística permite el análisis de respuestas nominales, y la regresión múltiple permite una o más respuestas continuas, lo que determina el método por usarse, dependiendo del análisis de datos o investigación que se esté realizando.

2.1 Visualización y análisis de información usando modelamiento de ecuaciones estructurales (SEM)

En la visualización de datos usando SEM, lo importante es mostrar diversas correlaciones entre las variables, dando varias perspectivas de la misma información, facilitando el manejo de filtros, lo que sugiere mayor o menor manejo de información en la visualización (TABLEAU, 2016), explorando los datos analizados principalmente mediante el sentido de la visión, usando las variables del modelo con la graficación, manipulando de forma dinámica atributos como forma de los componentes de la gráfica, orientación, color, textura, posición, tamaño y valor (Oliveira & Cardoso, 2014), como se realiza en la visualización y análisis dinámico mediante estructura en árbol (Yong & Yonghua, 2011), visualizándose información de una base de datos de forma dinámica.

Para la construcción de aplicaciones de visualización que usan SEM, se abordan técnicas como: brushing, panel matrix (Hyper-Slice y Hyperbox), iconografía, hierarchical display y Non-Cartesian display (Wong & Bergeron, 2015). Construir un componente para visualización y análisis requiere seleccionar entre los siguientes tres subcampos: visualización científica, visualización geográfica y visualización de información.

En este caso, la representación visual de información es la mitad de la decisión; la otra mitad es la interacción que se usará, de forma que permita modificaciones en la navegación y manipulación al momento de interactuar con la visualización para poder ver tendencias y patrones que no entrega una representación estática de forma tabular o una gráfica de líneas; se requiere que se tengan en cuenta elementos de percepción visual

humana y enfoque en áreas de interés como detail, 3d perspective, fisheye view o focus-plus-blur, que hacen más intuitivo el análisis y visualización de datos.

En el caso de uso de SEM para análisis y visualización, con frecuencia son realizadas visualizaciones de redes y grafos, lo que supone internamente que el análisis de datos se haga usando una matriz de adyacencia o matriz de diseño estructural entre los nodos, además de usar gráficas lineales con arcos y gráficas basadas en anillos (Andrews, 2016).

Es así como la visualización de información es una disciplina que ha crecido a la par con la computación, facilitando el renderizado de datos digitales (Wood et al., 2012), agrupándose las técnicas de visualización en seis clases: visualización de proyección geométrica, visualización basada en iconos, visualización orientada a pixels, visualización jerárquica, visualización basada en gráficos y visualización híbrida; estas a la vez se relacionan directamente con las técnicas de interacción de mapeo, proyección, filtrado, trazado y zoom para garantizar la interacción del usuario con la información con un alto nivel de usabilidad, como se muestra en la Figura 1 (Alvarado-Pérez & Bolaños-Ramírez, 2015).

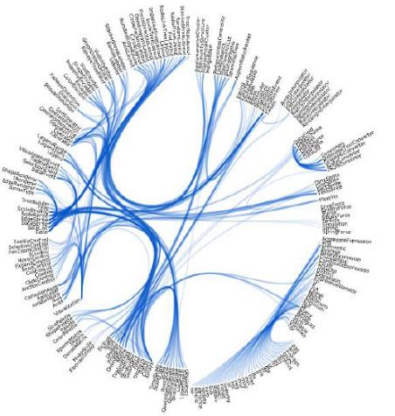

Figura 1. Visualización de información usando la librería Flare (PREFUSE, 2016).

Se logra que el uso de herramientas de visualización eficientes, fortalezca el pensamiento visual, ayudando al descubrimiento de información desconocida; de esta forma, la experiencia visual desempeña una parte importante en la acumulación de conocimiento y toma de decisiones, convirtiéndose en un material relevante de inspiración en el proceso cognitivo, reduciendo el costo del tiempo y complejidad en la adquisición de conocimiento en tiempo real (Fang, 2011).

En la actualidad existen algunos frameworks usados para hacer visualización y análisis de datos como son: Arbor.js, CaroDB, Chroma.js, Circos, Cola.js, ColorBrewer, ubism, Cytoscape, D3.js, Dance.js, Data.js, DataWrangler, Degr afa, Envision.js, Flare, GeoCommons, Gephi, Google Chart Tools, Google Fusion Tables, I Want Hue, JavaScript Infovis. Toolkit, kartograph, Leaflet, many Eyes, mapBox, Miso, modest Maps,

NVD3.js, nodeBox, OpenRefine, paper.js, Prefuse, Quadrigram, Sigma.js y Zingcharts (DATAVISUALIZATION, 2016).

De esta manera, dependiendo del tipo de visualización, se debe seleccionar el framework o herramienta adecuada, teniendo en cuenta la identificación de las variables (identificación del conjunto de las variables), determinación (estimación del cambio del modelo en el tiempo), visualización (forma de explicación de ideas), inferencia (inferencia causal entre los nodos que se relacionan), configuración (reordenar y cambiar características visuales) y localización (buscar rápidamente nueva información al realizar cambios de variables) (Pitchforth, 2013); todo esto permite que se seleccione uno u otro framework por su precisión y eficiencia respecto a cómo se desea realizar el análisis y visualización.

Igualmente, en algunas aplicaciones de visualización de información, esta puede ser realizada de forma pervasiva, permitiendo a varios usuarios por medio de diferentes dispositivos interactuar al mismo tiempo con un sistema multiusuario, como podría suceder con un par de personas que buscan una habitación de hotel desde lugares geográficamente diferentes, manipulando al mismo tiempo las variables: número de cuartos, lugar, precio, comodidades y demás, para lo cual el sistema debe transformar la vista en cada pantalla de los usuarios y al mismo tiempo coordinar las vistas de ambos usuarios, combinando sus selecciones y mostrando a ambos las selecciones de los dos (Craig, Huang, Chen, Wang & Zhang, 2015).

El anterior párrafo enfrenta uno de los principales problemas en la visualización de información, el cual es la latencia que se pueda producir, dependiendo de la cantidad de información que se está explorando de forma interactiva, lo que significa que se deba trabajar en el Modelo para Visualización de la Información (datos, acceso a los datos, tabla de datos, mapeo visual en primitivas y construcción de estructuras visuales (Carneiro, Teixeira, Aráujo, Santos & Junior, 2015). Este problema puede ser mejorado usando optimización en las búsquedas dinámicas y uso de un componente caché, con colas entre el acceso a los datos, el mapeo visual y el trazado en pantalla que permite la visualización final.

En la actualidad, se construyen aplicaciones web para visualización de datos, en las cuales se aplica el modelo de referencia de visualización de información, en el cual se hace análisis de información estructurada y no estructurada mediante la transformación de datos; haciendo *clustering* para construir un mapa de características, o haciendo transformación visual usando un grafo o un árbol en una fase de procesamiento global de los datos, para luego construir árboles o gráficos radiales resultado de un procesado detallado (Lirong, Mengjun & Jing, 2011), por transformación de los datos crudos a tablas de datos para hacer

mapeo visual y construir estructuras visuales que permitan al usuario visualizar los resultados.

Otra forma de optimizar los métodos de visualización jerárquica es usando el algoritmo Sunburst, el cual realiza una segmentación radial (node link) antes de hacer la visualización, enfocándose en maximizar solo las ramas que el usuario desea (Liu & Wang, 2015), aportando gran valor en el análisis, manejo de memoria y procesamiento para visualización de información.

En el análisis de datos en hábitos saludables, se recolectan los datos, se integran para analizarlos, y luego se realiza su visualización en tiempo real, de forma que los médicos pueden tener en tiempo real un diagnóstico sobre sus pacientes, requiriéndose el análisis de datos de la historia clínica electrónica del paciente y datos en tiempo real resultado de un tamizaje como temperatura, presión arterial y otros datos valiosos que puede usar el médico en una consulta (Ning, Wenxing & Siting, 2012), optimizándose el monitoreo en línea y diagnóstico de pacientes de forma proactiva.

Este tipo de soluciones también sirve para interacción presencial o en línea entre paciente y médico, en el cual se usan datos tomados de forma inmediata, identificándose información importante como valor (sí/no), severidad (normal, anormal), impacto del riesgo, perfil (información demográfica), signos vitales, historia del paciente, estilo de vida. Así el sistema puede ver factores o hábitos que afectan la salud y cómo aumentan el riesgo, mediante manejo de categorías y creación de vistas de aislamiento para analizar solo la información que interesa en el momento, creándose otra ventana para un análisis aparte (Bhaskaran, Kaduskar, Tallimani & Bhaumik, 2012).

Otro caso de uso de frameworks para visualización y análisis de datos médicos, incluye el uso de un diccionario ilustrado, el cual genera una gráfica en forma de rosas, dependiendo de la información médica multivariada que se obtenga, generando mensajes visuales en forma de rosa, usando mapeo de datos. De forma que el personal médico pueda hacer procesamiento de información automática y procesamiento inconsciente de información, de modo proactivo (Cai, Li, Zheng & Zhang, 2008), a través de la generación de cambios en el radio de las flores y su forma, su cantidad de pétalos y su color.

En este tipo de análisis, cuando se habla de datos orientados al tiempo, se hace uso de técnicas de parametrización e interacción de forma detallada, lo que demanda que los componentes del sistema de visualización de información (componente de visualización, componente de análisis y componente de eventos del usuario) estén totalmente integrados, mejorando el rendimiento del análisis de la

información visual, orientándose a la detección de eventos del usuario y representación de eventos del usuario (Aigner, Miksch, Wolfgang, Schumann & Tominski, 2009), lo que implica que el diseño de un sistema interactivo de visualización de información debe tener los siguientes componentes: cargado de datos, mapeo de datos, procesamiento de datos (Chen, 2013), generación de imágenes y control de eventos del usuario.

3. Conclusiones

La visualización y análisis multivariado de información tienen gran futuro y aplicación en todos los sectores productivos de la sociedad, mediante la implementación y desarrollo de investigaciones en ciencias computacionales en almacenamiento, comunicación, procesamiento, visualización y análisis de volúmenes de datos, lo que exige a los ingenieros de software, estadísticos y demás profesionales, trabajar de forma conjunta para lograr sistemas que utilicen modelos validados.

Se requiere gran atención cuando el volumen de datos analizados es significativo, lo que puede afectar el rendimiento y precisión del sistema, recomendándose que los componentes para interacción con el usuario, manejo del modelo SEM y visualización estén estrechamente acoplados, usando MVC y tecnologías de desarrollo que permitan la compatibilidad con las diferentes fuentes de datos y mejoren la experiencia de usuario y la usabilidad del sistema de análisis de datos.

Referencias

Aigner, W., Miksch, S., Wolfgang, M., Schumann, H. & Tominski, C. (2009). Visual Methods for Analyzing Time-Oriented Data. *IEEE Transactions on Visualization and Computer Graphics*, 1–13.

Alvarado-Pérez, J. C., Mariana, U. & Bolaños-Ramírez, H. (2015). Knowledge discovery in databases from a perspective of intelligent information visualization. *IEEE 2015 19th International Conference on Information Visualisation*, 1–7.

Andrews, K. (2016). *Information Visualisation*. Grraz.

Bhaskaran, P., Kaduskar, M., Tallimani, S. & Bhaumik, S. (2012). ForeTell : Facilitating doctor-patient conversation through interactive information visualization of risk prediction index. *IEEE 2012 International Conference on Bioinformatics and Biomedicine*, 615–618.

Cai, Z., Li, Y., Zheng, X. S. & Zhang, K. (2008). Applying Feature Integration

Theory to Glyph-based Information Visualization. *IEEE 2015 Pacific Visualization Symposium*, 99–103.

Carneiro, N., Teixeira, R., Aráujo, T., Santos, C. & Junior, J. (2015). A Concurrent Architecture Proposal for Information Visualization Pipeline. *IEEE 2015 19th International Conference on Information Visualisation A*. https://doi.org/10.1109/iV.2015.49

Chen, J. (2013). Research on interactive information visualization system in special academic discussion. *2012 Fourth International Symposium on Information Science and Engineering Research*, 59–63. https://doi.org/10.1109/ISISE.2012.22

Craig, P., Huang, X., Chen, H., Wang, X. & Zhang, S. (2015). Pervasive Information Visualization. *IEEE 2015 IEEE International Conference on Computer and Information Technology*, 2232–2233. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.330

DATAVISUALIZATION. (2016). Data Visualization Selected Tools. Retrieved June 18, 2018, from http://selection.datavisualization.ch/

Fang, S. (2011). Optimization for Information Visualization Based on Visual Thinking. *IEEE 2011 International Conference on Electronic & Mechanical Engineering and Information Technology Optimization*, 4243–4247.

Kan, Z., Xuefei, Z. & Jin, X. U. E. (2014). Evaluation of Quality and its Influence Factors of Human Settlement in the Metropolitan Periphery Area Based on Structural Equation Model, (January 2010), 1–5. https://doi.org/10.1109/ICMTMA.2014.70

Lirong, X., Mengjun, W. & Jing, F. (2011). A Visualization System for Web Retrieved Credit Information. *IEEE Seventh International Conference on Natural Computation*, 728–733.

Liu, C. & Wang, P. (2015). A Sunburst-Based Hierarchical Information Visualization Method and Its Application in Public Opinion Analysis.

IEEE 2015 8th International Conference on BioMedical Engineering and Informatics (BMEI 2015), (Bmei), 832–836.

Luse, A. (2014). Utilizing Structural Equation Modeling and Social Cognitive Career Theory to Identify Factors in Choice of IT as a Major, *14*(3), 1–19.

Ning, Z., Wenxing, H. & Siting, Z. (2012). A Solution for an Application of Information Visualization in Telemedicine. *IEEE The 7th International Conference on Computer Science & Education*, (Iccse), 407–411.

Oliveira, E. C. De & Cardoso, A. (2014). A Proposal for a Meta-Information Visualization using Treemap. *IEEE 2014 International Conference on Computational Science and Computational Intelligence*, 247–252. https://doi.org/10.1109/CSCI.2014.49

Pitchforth, J. (2013). An Evaluation of the Circles Information Visualization Tool for Presenting Bayesian Network Output. *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, 83–89. https://doi.org/10.1109/CIMSim.2013.22

PREFUSE. (2016). Dependency graph. Retrieved July 7, 2018, from http://flare.prefuse.org/launch/apps/dependency_graph

TABLEAU. (2016). *Visual Analysis Best Practices Simple Techniques for Making Every Data Visualization Useful and Beautiful*.

Templin, J. (2011). Univariate and Multivariate Statistical Distributions, 1–63.

Wong, P. C. & Bergeron, R. D. (2015). *30 Years of Multidimensional Multivariate Visualization*. Durham.

Wood, J., Isenberg, P., Isenberg, T., Dykes, J., Boukhelifa, N. & Slingsby, A. (2012). Sketchy Rendering for Information Visualization. *IEEE 2012 Transactions on Visualization and Computer Graphics*, *18*(12), 2749–2758.

Wuensch, K. L. (2016). *An Introduction to Multivariate Statistics*. Greenville.

Yong, L. & Yonghua, F. (2011). Information tree and information visualization. *IEEE 2011 2nd International Conference on Artificial Intelligence*, 1192–1195.

Fecha de recepción: 19 de julio de 2018.

Fecha de aprobación: 23 de julio de 2018.

Walter Hugo Arboleda Mazo
Corporación Universitaria Adventista
Investigador del Grupo de Investigación en Ingeniería Aplicada de la Facultad de Ingeniería de la Corporación Universitaria Adventista.
Ingeniería de Sistemas - Corporación Universitaria Adventista
Correo electrónico: warboleda@unac.edu.co
Leydy Johana Orozco Carvajal
Investigadora en las líneas de Gestión Tecnológica y Gestión del Conocimiento.
Administración de Empresas - Universidad Católica Luis Amigó
Correo electrónico: Leydy.orozcoca@amigo.edu.co