VISUALIZACIÓN DE GRANDES VOLÚMENES DE DATOS

Lic. Mercedes Vitturini[1] Lic. Karina Cenci[2] Mg. Silvia Castro[3]
[1] {mvitturi@criba.edu.ar} [2,3] {kmc,smc@cs.uns.edu.ar}
Departamento de Ciencias de la Computación
Universidad Nacional del Sur - Bahía Blanca

1. INTRODUCCIÓN

La disponibilidad de almacenamiento económico y el progreso tecnológico, han llevado a que se hayan creado inmensas bases de datos de negocios, de datos científicos, de datos meteorológicos entre otros tipos de datos. Ante el crecimiento tan vertiginoso en la cantidad de información de estas bases de datos y aún cuando las personas estén acostumbradas a interrogarlas, se hace prácticamente imposible para una persona la tarea de explorarlas para poder extraer conclusiones, tendencias y patrones. En este caso, sin duda los problemas de la consulta y la posterior exploración de las bases de datos son problemas clave. Con el objetivo de colaborar en la solución de los mismos se han desarrollado distintas herramientas de visualización.

Entre las primeras propuestas para la visualización de este tipo de información, surgen métodos interactivos basados en técnicas de browsing, de filtros y de facilidades para la construcción de consultas dinámicas que permitan aprender de los datos a través de múltiples consultas. Las propuestas de investigación más ambiciosas y recientes son las de data mining visual y están vinculadas con una nueva visión de la información en grandes bases de datos. Se pretende la búsqueda de nuevos conocimientos o profundización del discernimiento de conocimientos existentes, a través de un esfuerzo cooperativo entre el hombre y la computadora. Se basan en algoritmos de clustering guiados con técnicas de visualización interactivas para descubrir comportamientos y tendencias en los datos.

Informalmente, visualización es la transformación de datos o información en imágenes o pinturas. La visualización emplea el aparato sensitivo primario humano, que es la visión, tanto como todo el poder de procesamiento de la mente humana. El resultado debe ser un medio simple y efectivo para comunicar información voluminosa y compleja.

En este contexto, el objetivo de nuestro trabajo consiste en delinear criterios con el objetivo de obtener una visualización efectiva de grandes bases de datos en equipos de bajo costo.

En este trabajo presentamos una descripción de la investigación realizada sobre las tendencias y herramientas que se están utilizando para el discernimiento de grandes volúmenes de datos en Ciencias de la Computación. En la sección siguiente se detallan los conceptos fundamentales involucrados en visualización, visualización de información, data mining visual y cómo se relacionan a través de la representación gráfica de los datos. En la sección siguiente se presentan algunos ejemplos de visualización de bases de datos y se concluye con una descripción del trabajo a realizar.

2. VISUALIZACIÓN Y DATA MINING VISUAL

2.1 Visualización

Se puede definir la visualización como *la utilización de una computadora como soporte, interacción y representación visual* ***de los datos*** *para ampliar el conocimiento.* Esta definición de visualización focaliza tanto en el propósito de la visualización como en el de instrumento ó recurso. Hamming (1973) puntualizó que: *el propósito de la computación es el discernimiento y no los números.* En el caso de la visualización, podemos decir que *el propósito de la visualización es el discernimiento y no la imagen.* Las principales ventajas del discernimiento son el descubrimiento, la elaboración de decisiones y la posibilidad de explicar el comportamiento de los datos.

Muchas clases de información no tienen una representación física obvia y natural. La *visualización de información* permite visualizar espacios de información abstracta, tales como datos financieros, información de negocios, colecciones de documentos y concepciones abstractas que pueden también beneficiarse al ser presentadas en forma visual. El problema fundamental radica en mapear

abstracciones no espaciales en formas visuales efectivas, para lo cual es crucial descubrir nuevas metáforas visuales y entender qué tareas de análisis soportan.

Podemos definir entonces la *visualización de la información* como *la utilización de una computadora como soporte, interacción y representación visual de* ***datos abstractos*** *para ampliar el conocimiento.* Ante grandes volúmenes abstractos de información la meta es lograr *la cristalización del conocimiento,* es decir, permitir a los usuarios obtener la información que necesitan y hacer que ésta tenga sentido para que puedan lograrse las decisiones en un tiempo relativamente corto.

Como ejemplos de los objetivos de la visualización de información se pueden enunciar: mostrar tendencias en los datos, detectar discontinuidades en los mismos, identificar fácilmente máximos y mínimos, establecer límites, identificar agrupamiento en los datos, encontrar estructuras en información heterogénea y ver mucha información en una única pantalla pero al mismo tiempo ver un ítem de interés en este contexto, etc. La visualización de información apoya el proceso de producir modelos que puedan ser detectados y abstraídos; puede reducir la búsqueda de datos al agruparlos convenientemente o al relacionar la información visualmente, permite compactar información en un espacio reducido, permite búsquedas jerárquicas mediante la utilización de vistas generales para ubicar áreas de más detalle bajo demanda. La visualización permite la recuperación de modelos de datos y estos modelos sugieren esquemas a un nivel superior. La agregación de datos se revela a través de clustering o propiedades visuales comunes.

Esta exploración de la información requiere una interactividad óptima. La interacción con tiempos de respuesta de 1 segundo o menos agiliza el proceso de comprender los datos. Además permite al usuario explorar más posibilidades, dejando en la máquina el esfuerzo de procesamiento, en tanto el usuario observa qué pasa cuando se modifican los parámetros.

La interacción involucra la transformación de los datos a una forma visual en tanto el usuario maneja los *controles* para cambiar los parámetros de la cadena de transformaciones. Los controles pueden estar separados o integrados a la visualización. Algunos ejemplos de herramientas de control usadas son: botones; barras deslizantes sobre el alfabeto (*alphasliders*); ejes; barras deslizantes bidireccionales (*two-sided sliders*) y botones excluyentes. Cuando un usuario activa o modifica uno o más controles, el cambio debe reflejarse en tiempos interactivos.

2.2 Data Mining Visual

Data mining es la búsqueda de información valiosa en grandes volúmenes de datos. Es un esfuerzo cooperativo entre el hombre y la computadora.

Los hombres diseñan la bases de datos, describen problemas y objetivos conjuntos. Las computadoras clasifican los datos buscando patrones que responden a los objetivos. Data mining predictivo es una búsqueda de patrones fuertes en grandes conjuntos de datos que pueden generalizar decisiones futuras precisas.

Data mining (también denominado Descubrimiento del Conocimiento en Bases de Datos) es el descubrimiento eficiente de patrones previamente desconocidos en bases de datos.

A medida que es ampliamente reconocido que los datos son un recurso valioso para cualquier organización, extraer información a partir de los mismos es a menudo un problema dificultoso pero vital. Cada tarea típica de data mining requiere una solución a medida que depende del carácter y de la cantidad de los datos. Cuando se trabaja en algoritmos de data mining es deseable obtener propiedades tales como: descubrir patrones en grandes bases de datos, en vez de simplemente verificar que el patrón existe; tener una propiedad de integridad que garantice que todos los patrones de ciertos tipos hayan sido descubiertos; tener una ejecución rápida y un escalamiento cercano al lineal en bases de datos reales muy grandes (del orden de los terabytes)

A través de los años se han propuesto diferentes métodos para la clasificación, siendo los árboles de decisión particularmente convenientes para el data mining, ya que pueden ser construidos relativamente rápido comparados con otros métodos además de ser fáciles para interpretar. Los árboles pueden ser convertidos en sentencias SQL que pueden ser utilizadas para el acceso eficiente de las

bases de datos. Los problemas se organizan en dos categorías generales: predicciones y descubrimiento de conocimiento

El aplicar una visualización a los datos en una gran base de datos sin un paso previo de preprocesamiento conduce a menudo a resultados incomprensibles. La *visualización de información necesita de las técnicas de data mining* para poder extraer y visualizar sólo los resultados que se deseen. Este descubrimiento de información a partir de los datos puede ser supervisado o no: en el primer caso, los usuarios tienen una meta particular en mente, en tanto que en el otro caso los usuarios llevan a cabo una búsqueda con el propósito de encontrar patrones de interés.

Data mining necesita del uso de técnicas de visualización interactivas que permitan al usuario cambiar rápida y fácilmente tanto el tipo de información mostrada como el método de visualización. Las visualizaciones son particularmente útiles para descubrir fenómenos que se mantienen para un subconjunto relativamente pequeño de datos. La ventaja de usar visualización es que el usuario no necesita conocer el tipo de fenómeno que está buscando para descubrir algo inusual o interesante.

Podemos concluir entonces que un sistema de data mining visual está basado en los siguientes principios: simplicidad, autonomía del usuario, confiabilidad, reusabilidad, disponibilidad y seguridad. Debe ser sintácticamente simple de ser utilizado; simple no significa que sea trivial o no poderoso. Un sistema de data mining visual genuino no debe imponer conocimiento en los usuarios, sino que debe guiarlos a través del proceso de entendimiento para que estos elaboren las conclusiones. Las personas deberían estudiar las abstracciones de la visualización y obtener un mayor nivel de compresión en vez de aceptar una decisión automática.

4. TRABAJO RELACIONADO

Luego de haber presentado los conceptos teóricos básicos en visualización de información aplicada a bases de datos, presentaremos los casos de estudio que se investigaron y que se consideraron significativos en lo referente a la evolución de la visualización de la información aplicada a bases de datos con objetos de varias dimensiones:

- VisDB [8], esta aplicación presenta un método para visualizar un gran conjunto de datos; cada dato se representa con un pixel y su posición y color dependerán de cuan relacionado se encuentra con la consulta original. Favorece el estudio de grandes cantidades de datos de los que se quiere descubrir su comportamiento. La posibilidad de realizar consultas y modificaciones a la misma en forma interactiva permite que el usuario obtenga realimentación y descubra de esta manera comportamientos interesantes en las respuestas.

- Filmfinder[12] es un buscador interactivo de películas. Los datos se presentan sobre un plano comparativo con posibilidades de elaborar las consultas en tiempo real a través de herramientas gráficas. Permite realizar una profundización de la semántica del entorno cuando se necesita aumentar el detalle. Las consultas se realizan exitosamente para pequeñas bases de datos pero son más lentas a medida que el tamaño de la base de datos aumenta.

Entre los desarrollos que incorporan más herramientas de data mining se pueden citar [9], [10], [11]. Los desarrollos [9] y [10] utilizan técnicas de clustering con visualización y participación interactiva del usuario, para descubrir agrupamiento entre los datos. En [9] se presenta una aproximación de clustering interactiva rápida para lograr clustering óptimo. En [10] se combinan algoritmos de clustering con técnicas de visualización orientadas a píxel y con representación icónica de los datos; los usuarios deben especificar los separadores de clustering en la separación. El [11] presenta una novedosa visualización basada en formatos de 3D que permite comparar el contenido de documentos de texto.

5. TRABAJO FUTURO

La explosión en cantidad, tamaño y disponibilidad de las fuentes de información ha despertado el interés en data mining, que permite exploración y descubrimiento interactivo de conocimiento y relaciones subyacentes en esa gran cantidad de datos. La visualización amplifica este proceso de

análisis. Sin duda, el acoplamiento de visualización, data mining y técnicas de análisis de información, constituyen un nuevo y poderoso paradigma para la exploración y el descubrimiento.

Hasta el momento no hay una integración sistemática de las técnicas de Visualización de Información en el diseño de sistemas que almacenan una gran cantidad de información. Es necesario encontrar los elementos y delinear los criterios que le permitan a los diseñadores seleccionar las herramientas apropiadas para cubrir los requerimientos.

Los estudios realizados constituyen el punto de partida para lograr estas metas con el objetivo de obtener una visualización efectiva de grandes bases de datos.

Hasta el momento:

- Se ha iniciado el estudio de los principios básicos de la visualización.
- Se ha realizado una revisión de los distintos métodos y herramientas de visualización de grandes bases de datos; además se están analizando exhaustivamente los distintos métodos existentes con fines de comparación.

Se debe dilucidar qué características conducen a resultados más exactos, a una mejor productividad y a un mejor entendimiento de los patrones y comportamiento generales y/o particulares de los datos. Concretamente, se pretende utilizar los resultados obtenidos en una aplicación real con datos académicos de los alumnos de una universidad. El objetivo es descubrir patrones de alumnos que logran finalizar sus estudios, los abandonan en primer instancia o después de haber aprobado los primeros años, etc.. Cabe destacar que, debido a las características de la aplicación, el objetivo principal es lograr una visualización efectiva en equipos de bajo costo. Inicialmente se analizarán, en este contexto, los métodos existentes; así, se pretenden delinear criterios para obtener una visualización adecuada que muestre una aproximación inicial del comportamiento de los datos, luego, y mediante técnicas apropiadas de interacción, se debe permitir la participación del usuario para lograr una visualización efectiva.

6. BIBLIOGRAFÍA

[1] McCormick, B. H., DeFanti, T.A., Brown, M.D., *Visualization in Scientific Computing*, Report of the NFS Advisory Panel on Graphics, Image Processing and Workstations, 1987.

[2] Schroeder, W., Martin, K., Lorensen, B., *The Visualization Toolkit An Object-Oriented Approach to 3D Graphics*, Prentice Hall, 1998.

[3] Card, S., Mackinlay, J., Shneiderman, B., Eds., *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Pub., 1999.

[4] Chung Wong, P., *Visual Data Mining*, IEEE Computer Graphics and Applications, pp. 20-21, September-October 1999.

[5] Agrawal, R., Shafer, J., *Parallel Mining of Association Rules.*

[6] Grady, N., Schryver, J., Leuze, M., *Mining for Personel Profiles.*

[7] Agrawal, R., Mehta, M., Shafer, J. y otros. *The Quest Data Mining System.*

[8] Keim, D., Kriegel, H., *VisDB: Database Exploration using Multidimensional Visualization*, IEEE Computer Graphics and Applications,.pp. 40-49, Sept. 1994.

[9] Ribarsky, W., Katz, J., y otros, *Discovery Visualization using Fast Clustering*, IEEE Computer Graphics and Applications, pp. 32-39, September-October 1999.

[10] Hinneburg, A., Keim, D., Wawryniuk, M., *HD-Eye: Visual Mining of High-Dimensional Data*, IEEE Computer Graphics and Applications, pp. 22-31, September-October 1999.

[11] Rohrer, R., Ebert, D., *A Shaped-based Visual Interface for Text Retrieval*, IEEE Computer Graphics and Applications, pp. 40-45, September-October 1999.

[12] Ahlberg, Ch., Shneiderman, B., *Visual Information Seeking: Tight Coupling of Dymanic Queries Filters with Starfield Displays*, Proceedings of CHI'94, ACM Conference on Human Factors in Computing Systems, pp. 313-317, New York.