Mapas visuales para el análisis interactivo de datos multidimensionales

Ana Maria Cuadros

Ana Maria Cuadros Valdivia es magíster en Ciencias de la Computación, en la especialidad de Computación Gráfica, por la Universidade de São Paulo (Brasil) e ingeniera informática por la Universidad Católica San Pablo (Arequipa), universidad donde se desempeña como docente. Actualmente es candidata a doctora en Ciencias de la Computación por la Universidad Nacional de San Agustín (Arequipa)

Mapas visuales para el análisis interactivo de datos multidimensionales

Resumen

Las técnicas de visualización de información son una herramienta preponderante en el análisis de datos estructurados y no estructurados con variada dimensionalidad. El objetivo de este trabajo es introducir una nueva técnica que mediante mapas visuales ayude al usuario en el análisis de grandes conjuntos de datos multimedia. Los mapas emplean técnicas de posicionamiento de puntos. El punto de partida son las similitudes, que sirven para construir espacios visuales que permitan la exploración visual y la interacción con los diferentes tipos de datos en un solo ambiente visual.

Palabras clave: minería visual de datos, minería de datos multimedia, análisis de datos, visualización, fusión de datos multimodales, mapas visuales

Los cambios en las tecnologías de información y comunicación han creado en la última década nuevos hábitos de trabajo, nuevas formas de almacenamiento de información y nuevos medios de comunicación y difusión. Un ejemplo característico son las colecciones de documentos con contenido similar almacenados en diferentes formatos (HTML, Microsoft Office, Open Office, PDF, entre otros), los cuales comprenden texto, imágenes, sonido y video: si en un inicio la información generalmente sólo se almacenaba en formato texto, ahora se almacena en modernos documentos multimedia que —contrariamente al texto— contienen una serie de formatos válidos en su contenido (Magalhães e Iria, 2009).

Ante el cúmulo de aplicaciones multimedia, y ante la diversa naturaleza y complejidad de sus datos, existe la necesidad de crear métodos y herramientas que descubran las relaciones existentes entre aquéllos, y que apoyen efectivamente en las tareas de análisis para permitir al usuario extraer información y conocimiento relevante para tomar decisiones. Pero ¿cómo descubrir ese conocimiento de forma rápida y efectiva en tiempo real?

Un área que viene direccionando todo su esfuerzo en resolver este desafío es la minería de datos multimedia. Su propósito es procesar información multimedia de alto nivel —sola o en combinación con otros tipos de datos— para encontrar patrones útiles que sirvan para la toma de decisiones (Petrushin y Khan, 2007). Sin embargo, las técnicas utilizadas para extraer conocimiento generalmente se encuentran con los siguientes inconvenientes: (1) dificultad para extraer datos relevantes y útiles a partir de datos heterogéneos (combinación de texto, imágenes, video y sonido en una única instancia); (2) cantidad de atributos asociados a cada instancia, lo que ocasiona una elevada dimensionalidad en los datos; (3) conocimiento a priori del comportamiento de los datos, ya que los algoritmos de minería no trabajan efectivamente de forma no supervisada; y (4) los algoritmos utilizados consiguen extraer gran cantidad de patrones, lo que dificulta muchas veces la interpretación de los resultados.

Frente a estos problemas una posible solución es aplicar técnicas de visualización de información durante la ejecución de algoritmos de minería para presentar resultados parciales y finales, lo cual podría ayudar a los usuarios en la interpretación de los resultados (Ankerst, Keim, y Kriegel, 1996; Card, Mackinlay, y Shneiderman, 1999)

En este contexto, una de las técnicas de visualización que ha despertado el interés de la comunidad científica es la de *mapas de documentos* (Paulovich, Oliveira, y Minghim, 2007). Un mapa de documentos es un espacio de información visual que permite la navegación del usuario; en el mapa aparecen reflejadas las relaciones de similitud entre los datos por medio de diferentes representaciones geométricas. Los mapas pueden soportar una va-

riedad de tareas exploratorias, de modo que conectan a los usuarios con sus respectivos mapas cognitivos del espacio de información visual (Chen, 2006; Paulovich y Minghim, 2006). De este modo facilitan el acceso al conocimiento inmerso en colecciones grandes y complejas de documentos, y mejoran la efectividad en la toma de decisiones (Becks, Seeling, y Minkenberg, 2005).

En la literatura existen diversas técnicas para realizar mapeo de documentos. En este trabajo nos centraremos en aquéllas basadas en proyecciones y construcción de filogenias. Las técnicas basadas en proyecciones permiten identificar zonas de similitud entre las instancias de datos, lo cual torna más sencilla la interpretación, comprensión y extracción de información (Eler et al., 2009; Paulovich, Nonato, Minghim, y Levkowitz, 2008; Paulovich, et al., 2007). Las técnicas basadas en construcción de filogenias ubican con un alto grado de precisión documentos en vecindades adecuadas, lo cual facilita la comprensión entre documentos similares mediante el uso de jerarquías (Cuadros, Paulovich, Minghim, y Telles, 2007).

Basada en estos trabajos propondré el uso de mapas de documentos para generar clasificación no supervisada de documentos multimedia multidimensionales. En concreto pretendo integrar imágenes y texto para ser analizados y explorados de manera uniforme, como si se tratara de un único tipo de dato multidimensional. Al mismo tiempo, intentaré mostrar que integrar ambos tipos de datos en un simple ambiente de exploración mejora la capacidad de revelar información interesante para los usuarios.

Diversas investigaciones han trabajado métodos para la exploración visual de conjuntos de datos; muchas han empleado también mapas de documentos, pero aplicados a un solo tipo de datos. En esta investigación me concentraré en aquéllas basadas en contenido, es decir, en visualizaciones destinadas a ayudar a ubicar documentos correlacionados, de tal manera que el usuario no tenga que leer muchos para tener una buena noción de los conceptos abordados en una determinada colección de documentos.

Los enfoques más conocidos que interactúan con mapas de documentos textuales son los basados en proyecciones multidimensionales. Dos técnicas clásicas para visualizar textos basados en contenido son Gallaxy (Wise, 1999) —implementada ahora en IN-SPIRE (PNNL, 2010), un sistema de visualización de textos comercial— e InfoSky (Andrews et al., 2002). Tanto Gallaxy como InfoSky permiten visualizar conjuntos de documentos textuales en espacios bidimensionales procurando colocar textos muy próximos de acuerdo con una medida de similitud. Para ubicar grupos de textos en determinadas regiones de la visualización se utilizan técnicas de agrupamiento y proyección.

Otra herramienta, en este contexto, es Projection Explorer (PEx), propuesta por Paulovich (Paulovich, et al., 2007). Para construir mapas visuales de conjuntos de documentos textuales basados en contenido PEx se hace uso de diversas técnicas rápidas de proyección multimensional y de posicionamiento de puntos. Recientemente Eler (Eler, et al., 2009) modificó PEx para explorar conjuntos de imágenes basados en similitud por contenido, lo que dio origen a PEx-Image. PEx y PEx-Image se basan en un enfoque de grafos, donde los nodos representan textos multidimensionales, mientras que las aristas, las relaciones entre ellos. La representación se realiza en un espacio bidimensional y permite visualizar diferentes relaciones de vecindad tanto en el espacio bidimensional como en el multidimensional. La coloración de nodos se da de acuerdo con la frecuencia de palabras o grupos de palabras en los documentos; al mismo tiempo, la creación de etiquetas *(tags)* permite identificar los principales temas en un grupo de documentos seleccionados.

Ambas técnicas se probaron con éxito en trabajos previos (Cuadros, et al., 2007; Eler, et al., 2009; Paulovich, et al., 2008; Paulovich, et al., 2007). Con ellas se construyeron mapas visuales para diversos conjuntos de instancias de datos. En el caso de instancias de tipo texto se realizaron pruebas principalmente sobre noticias breves de periódicos, artículos científicos, patentes, *newsgroups,* entre otros. Para el caso de datos de tipo imágenes se realizaron pruebas sobre conjuntos de imágenes médicas, fotografías y secuencias de proteínas.

A partir de estos trabajos adaptaré y extenderé las herramientas PEx y PEx-Image para generar una propuesta original de clasificación no supervisada de documentos multimedia multidimensionales.

Proyecciones multidimensionales vía posicionamiento de puntos

Una manera de manipular la elevada dimensionalidad de los datos es reducir el número de dimensiones, por lo que se pueden aplicar estrategias que funcionan bien con bajas dimensiones. Las técnicas de proyecciones multidimensionales son un ejemplo. Típicamente, una técnica de proyección de datos multidimensionales mapea datos m-dimensionales en un espacio p-dimensional, con $p = \{1, 2, 3\}$, preservando alguna información sobre las relaciones de distancia entre las instancias de datos, de forma que revelen al máximo las estructuras del conjunto de datos. De esta manera se pueden crear representaciones gráficas para tomar ventaja de la percepción visual, y reconocer patrones basados en similitud, tales como agrupaciones de elementos.

Para este propósito se crearon diversas técnicas de reducción de dimensionalidad. Entre ellas están las clásicas Self-Organizing Map (SOM), Principal Component Analysis (PCA), Latent Semantic Index (LSI) y Multidimensional Scaling (MDS). Sin embargo, también se crearon técnicas *rápidas* de proyección multidimensional como Projection by Clustering (ProjClus) (Paulovich y Minghim, 2006) y Least-Square Projection (LSP) (Paulovich, et al., 2008). Ambos tipos de técnicas tuvieron éxito construyendo mapas de documentos textuales y son herramientas valiosas para ayudar a los usuarios a extraer información relevante de un conjunto de documentos.

Aunque cada técnica de proyección adopta su propio formalismo matemático para mapear datos de espacios de alta dimensionalidad a unos de menor dimensionalidad (de 1, 2 ó 3 dimensiones), todas tienen como objetivo común generar grupos densos de instancias multidimensionales semejantes y mantenerlos lo suficientemente separados cuando se trata de grupos disímiles.

La principal ventaja de estas técnicas es su poca complejidad computacional y, dependiendo del tipo de proyección, son adecuadas en procesos incrementales (Telles, R, y Paulovich, 2005). Su desventaja es que no preservan las relaciones de los datos definidas en su espacio original.

Reconstrucción filogenética: Neighbor Joining-Tree

La construcción de un árbol Neighbor Joining (NJ) (Saitou y Nei, 1987) parte de la suposición de que existen *n* objetos o instancias que se deben relacionar a través de un árbol filogenético; al mismo tiempo, presupone que las distancias D_{ij} para cada par de objetos *(i, j)* son conocidas.

NJ empieza con un árbol estrella con *n* nodos conectados a un simple nodo interno. En cada paso, el algoritmo (1) selecciona un par de nodos *(i, j)* que tienen el menor valor en el cálculo de la suma de las distancias entre las ramas $S_{i,j}$; (2) adiciona un nodo *x* al árbol, con los nodos *i* y *j* como hijos, y es conectado al ancestro común de *i* y *j;* (3) evalúa la longitud de las ramas L_{ix} y L_{jx}; y (4) reemplaza *i* y *j* por *x* en la matriz de distancia. El ciclo se repite hasta que el número de objetos sea igual a 3.

El presente trabajo se basa en la técnica Neighbor Joining-Tree (NJ-T) (Cuadros, et al., 2007), que si bien emplea el razonamiento adoptado por el método NJ, lo aplica a conjuntos de datos multidimensionales, como imágenes y textos. El concepto de ancestra-

lidad es reemplazado por nodos virtuales con disimilitud combinada. El árbol se crea observando la propiedad de que los nodos similares se deben asignar a la misma rama, en una estrategia *bottom-up*. Por consiguiente, un nodo se definirá como ancestro de otro cuando ambos tengan un contenido similar.

El árbol generado no tiene raíz, y su poca profundidad permite el uso racional del espacio de visualización; al mismo tiempo, su útil interpretación jerárquica deja apreciar las relaciones de similitud local y global. La principal ventaja de la técnica NJ-T es que ubica documentos en vecindades adecuadas con un alto grado de precisión, lo cual facilita la comprensión entre documentos similares haciendo uso de jerarquías. Su principal desventaja, sin embargo, es su elevado costo computacional.

A partir de las técnicas analizadas en esta sección desarrollé una propuesta propia para construir mapas visuales de documentos multimedia.

Método

Parto del supuesto de que los documentos multimedia son una rica fuente de información producida y analizada por los seres humanos. Los usuarios que lidian con este tipo de información probablemente no saben a priori el contenido de los documentos, de modo que no pueden guiarse por la relevancia de la información para encontrar la que buscan. Con el fin de aliviar esta tarea propongo la construcción de mapas visuales de documentos que permitan capturar las relaciones de proximidad de documentos multimedia. Esto ofrecerá al usuario una visión global del conocimiento presente en un conjunto de documentos y le permitirá tener una visión local con varios niveles de detalle, lo cual le hará más fácil la formulación de consultas.

El proceso general para construir, analizar, explorar e interactuar con los mapas de documentos multimedia basados en contenido se puede ver en la figura 1. Este proceso está compuesto por siete etapas:

1. Extracción de características de textos.
2. Extracción de características de imágenes.
3. Fusión en un solo vector de características.
4. Cálculo de la matriz de similitud mediante la aplicación de alguna métrica para establecer un criterio de similitud entre los documentos multimedia.
5. Construcción del árbol usando el algoritmo de filogenia NJ-T.

6. Posicionamiento del conjunto de vértices en la superficie de visualización usando un *layout* de árboles sin raíz.
7. Exploración e interacción con el árbol generado.

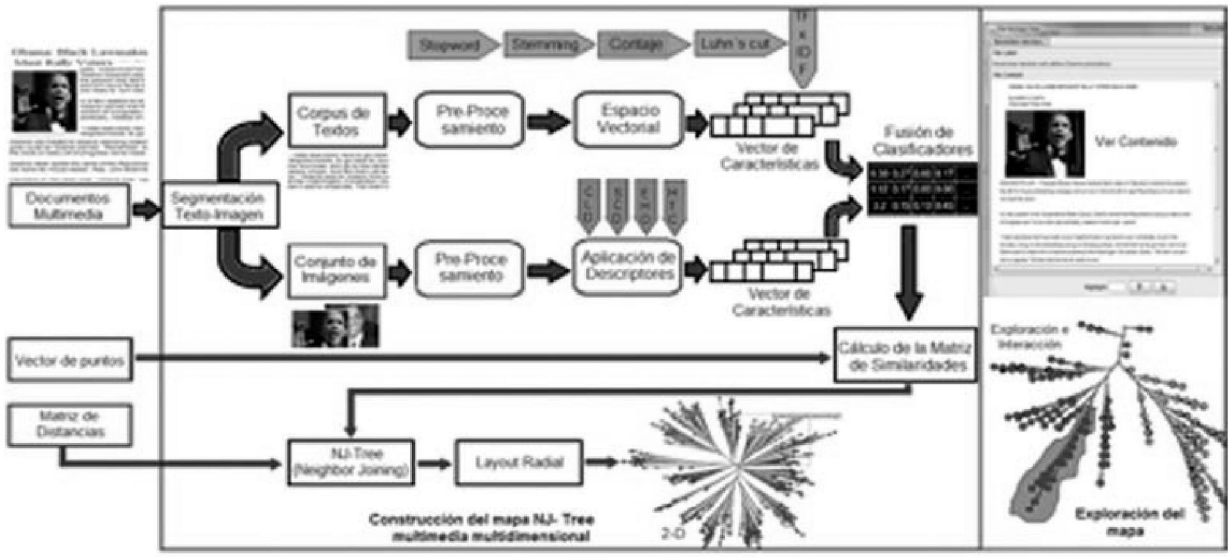

Figura 1
Proceso de construcción de mapas interactivos de documentos multimedia mediante la técnica NJ-T y la herramienta PEx.

El sistema recibe como datos de entrada un conjunto de documentos multimedia *d,* cuyo contenido consiste en datos textuales y en una única imagen.

El primer paso es convertir la parte textual en un vector espacial de *n* dimensiones $T_d = (t_1,t_2,\ldots t_n)$, cuyas coordenadas reflejarán la frecuencia de cada término presente en el texto. Para crear un vector representativo en un espacio multidimensional se siguen los siguientes pasos: (1) remoción de términos no representativos de los documentos *(stopwords);* (2) identificación de radicales y eliminación de sufijos y prefijos *(stemming)* de los términos mediante el algoritmo de Porter (1980); (3) conteo de las palabras dentro de cada documento para determinar sus frecuencias de ocurrencia, para lo cual se aplican los cortes superiores e inferiores de Luhn (Luhn's Cut) (Luhn, 1968), que eliminan palabras frecuentes o muy raras; finalmente (4) atribuir un valor (peso) de acuerdo con la importancia de cada término en el texto, de acuerdo con el TF[1]x IDF[2]. El resultado de este proceso es un vector que representa al texto de un documento de *n*-dimensiones (términos).

1. Term Frequency.
2. Inverse Document Frequency.

El segundo paso consiste en extraer los descriptores de una imagen, consolidados en un vector de características *m*-dimensionales $I_d = (i_1, i_2,\ldots i_m)$. Se emplean cuatro descriptores visuales para extraer las características de la imagen (Manjutah, Ohm, Vasudevan, y Yamada, 2001): Color Layout (CLD), Scalable Color (SCD), Edge Histogram (EHD) y Homogeneous Texture (HTD). Se utiliza CLD para capturar la distribución espacial del color en una imagen. El proceso de extracción de características consiste en cuatro pasos: (1) partición de la imagen; (2) detección del color dominante; (3) transformada discreta del coseno (TDC); y (4) cuantización de los coeficientes TDC con el escaneado zigzag. SCD es un descriptor que consiste en un histograma de color de 256 bins en el espacio de color HSV[3], codificado con una medida de Haar. EHD calcula la distribución espacial de los bordes. Se detectan cuatro direcciones en los bordes (0°, 45°, 90° y 135°) y bordes no direccionales. El EHD es un descriptor que nos facilita información sobre el tipo de contornos o bordes que aparecen en la imagen, lo que permite diferenciar si son bordes horizontales, verticales, oblicuos o aleatorios. HTD filtra la imagen con un banco de orientación y una escala de filtros sintonizados para obtener 30 filtros de Gabor.

Los cuatro descriptores se unen en un único vector. Si D_{CLD}, D_{SCD}, D_{EHD}, D_{HTD} son los cuatros descriptores mencionados anteriormente, el descriptor combinado sería igual a:

$$D_{combinado} = [D_{CLD} | D_{SCD} | D_{EHD}\, D | D_{HTD}]$$

La razón para optar por estos descriptores es que ya están normalizados. Todas las características deberían tener más o menos el mismo valor numérico para evitar los efectos de escala. En este caso los descriptores son escalados por valores enteros de magnitud equivalente. Para evaluar la combinación de cada uno de los descriptores en un único vector se utilizó la técnica Super Vector Machine (Vapnik, 1995).

Una vez que se obtienen los vectores de texto y de imágenes por separado, el tercer paso es normalizar y concatenar los vectores que corresponden a un mismo documento multimedia de *n* x *m* dimensiones:

$$D_d = [t_1, t_2,\ldots t_n | i_1, i_2,\ldots i_m]$$

Para normalizar los vectores se atribuyen pesos de acuerdo con una probabilidad; estos pesos se basan en reglas heurísticas, tal como hicieron Kalva y sus colaboradores (Kalva, Enembreck, y Koerich, 2007). El resultado de este proceso es una matriz en la que cada línea (vector) representa un documento, y las columnas (dimensiones) representan las características.

3. Hue, Saturation, Value.

El cuarto paso es construir una matriz de similitud M (d x d), donde M_{ij} es la distancia entre los documentos multimedia i y j. La medida usada para definir la distancia entre los documentos es la métrica basada en el coseno.

Una vez construida la matriz de similitud, el quinto paso es construir un árbol filogenético usando la técnica NJ-T, que genera como salida un árbol sin raíz, como ya se explicó. La estructura del árbol se exhibe en una superficie de visualización por medio de la aplicación de un algoritmo de diseño de árboles. Los nodos hoja representan los documentos multimedia; los nodos internos representan nodos hipotéticos; las ramas reflejan las relaciones de similitud entre los nodos.

El sexto paso es diseñar el *layout* radial del árbol (Bachmaier, Brandes, y Schlieper, 2005). Como resultado inicial es posible percibir visualmente la estructura global del árbol y la formación de los grupos potenciales del conjunto de documentos.

En el paso final el esparcimiento se puede realizar de los nodos, de manera que el usuario puede navegar e interactuar con los detalles de la información contenida en el árbol. En términos de funcionalidades que ayudan a los usuarios en las tareas de interacción con los árboles y el mapa en general, el analista puede recurrir a las tareas de ampliación y reducción de jerarquías del árbol, así como al desplazamiento de las ramas, para facilitar al máximo el proceso de exploración cuando ocurra una sobreposición de nodos y ramas. En ese sentido, para identificar las relaciones de vecindad entre nodos se implementó la funcionalidad Neighbor Depth. Si se diera un clic simple sobre un nodo, sus vecinos serían identificados de acuerdo con una trayectoria que sigue el nodo en el proceso de formación del árbol.

De otro lado, se adaptaron funcionalidades ya implementadas en la herramienta PEx (Paulovich, et al., 2007) para su funcionamiento en los árboles, tales como selección de ramas, visualización del contenido de los documentos y coordinación de múltiples visiones. La coordinación implica que el usuario pueda seleccionar un documento o un grupo de documentos en una vista resaltando los documentos correspondientes o los relacionados en otra vista coordinada.

Resultados y discusión

Para evaluar la propuesta se realizaron varias pruebas sobre diversos conjuntos de datos multimedia. Presentaré particularmente los resultados obtenidos al aplicar dos conjun-

tos de datos multidimensionales: noticias de periódicos extraídas de la web e imágenes médicas. El cuadro 1 presenta un resumen de esto.

Cuadro 1
Conjunto de datos utilizados para mostrar la validez de la técnica de mapeo de datos multimedia

Conjunto de datos	Tipo	Tipos de datos	N.º de datos
NEWS	Noticias breves	Imagen, texto	74
Imágenes médicas	Tomografías	Imagen	270

Para validar el mapeo de datos multimedia primero experimentamos con el tipo de datos imágenes. El objetivo es probar la eficiencia de los descriptores de características de las imágenes. La figura 2 muestra la construcción del mapa multidimesional aplicado a imágenes médicas. El vector de características se obtuvo al normalizar los valores generados por cada uno de estos descriptores: *color layout, edge histogram, scalable color* y *homogeneus texture.* La métrica de distancia utilizada es la del coseno. Las 270 imágenes utilizadas se distribuyeron en seis clases, cada una de las cuales se identificó en el mapa con diferentes colores.

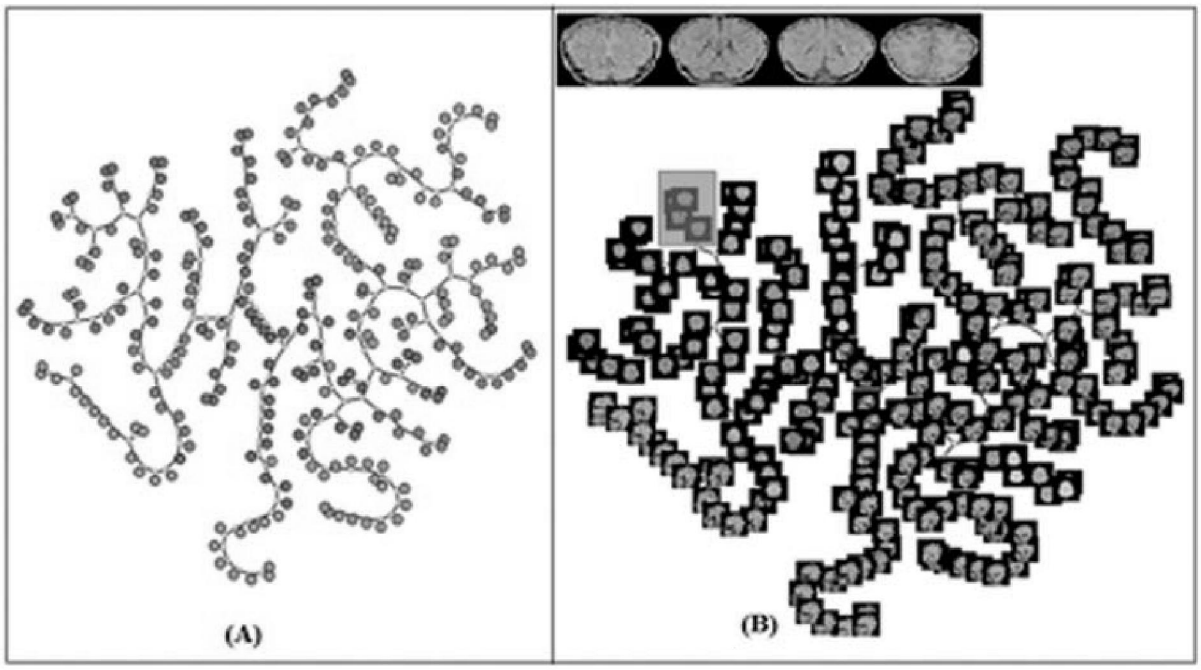

Figura 2
(A) Proyección 2D de la base de datos de imágenes médicas. El *layout* se generó con la técnica NJ-T usando la distancia del coseno. Los colores representan las clases de imágenes; (B) visualización de los vecinos más cercanos (imágenes similares) de una selección de imágenes realizada por el usuario.

De acuerdo con el mapa de imágenes de la figura 2, podemos inferir que las imágenes son separadas en ramas de acuerdo con la similitud en su contenido. De este modo, muchas imágenes de una determinada clase se han mapeado dentro de la misma rama o en ramas cercanas. Además, las imágenes con un alto grado de similitud son colocadas en los extremos de las ramas. La figura 2B muestra las imágenes en miniatura como marcadores gráficos implementados en la herramienta PEx-Image (Eler, et al., 2009). Podemos concluir que los descriptores de imágenes propuestos conjuntamente con la técnica de posicionamiento de puntos NJ-T permiten distribuir los nodos en el espacio de visualización, separando las imágenes similares conforme a la preclasificación realizada a priori del conjunto de imágenes médicas.

Para analizar el comportamiento de nuestra técnica con respecto a la fusión de datos multimedia —imágenes y textos— se empleó el conjunto de datos NEWS, compuesto por noticias seleccionadas de agencias como Associated Press, Reuteurs, BBC y CNN. Los artículos fueron publicados entre los meses de agosto y setiembre del año 2010. Los documentos multimedia consisten en una breve nota con la noticia y una imagen representativa que describe su contenido. El conjunto de datos consiste en 74 noticias divididas en 6 clases: noticias del arresto de Lindsay Lohan; la liberación de Sarah Shourd en Irán; el descubrimiento de una nueva especie animal; la invasión de especies animales foráneas en territorio europeo; el rescate de mineros atrapados en Chile; y las elecciones del nuevo Congreso estadounidense que se realizarían en noviembre. Los nodos en el mapa son coloreados de acuerdo con la clase a que pertenecen. Este conjunto de datos se usó para medir la efectividad de la fusión de las características extraídas de las imágenes y texto en un único vector.

Para la validación previamente se generaron los mapas correspondientes a un solo modo de datos multimedia por separado. La figura 3 muestra la construcción del mapa con el uso del contenido textual de las noticias.

Como se aprecia, la técnica NJ-T aplicada a vectores de características textuales consiguió ubicar las seis clases de tópicos en diferentes ramas. Para analizar si las ramas efectivamente contienen temas de asuntos similares se aplicó una de las funcionalidades de la herramienta PEx: la posibilidad de crear rótulos por grupos por medio del algoritmo de agrupamiento K-Means (Anderberg, 1973; MacQueen, 1967). Se generaron 8 grupos, de los cuales sólo 4 coinciden con los subárboles generados por NJ-T. Estos corresponden a los sucesos relativos a Lindsay Lohan, a las elecciones legislativas en EE. UU., al rescate de los mineros y a la liberación de la americana Sarah Shourd. Los dos restantes son divididos en subtemas específicos. Por ejemplo, en el caso del tópico de invasión de

especies foráneas, no sólo afecta al hábitat animal sino también al vegetal, por lo que las ramas se subdividen en subramas.

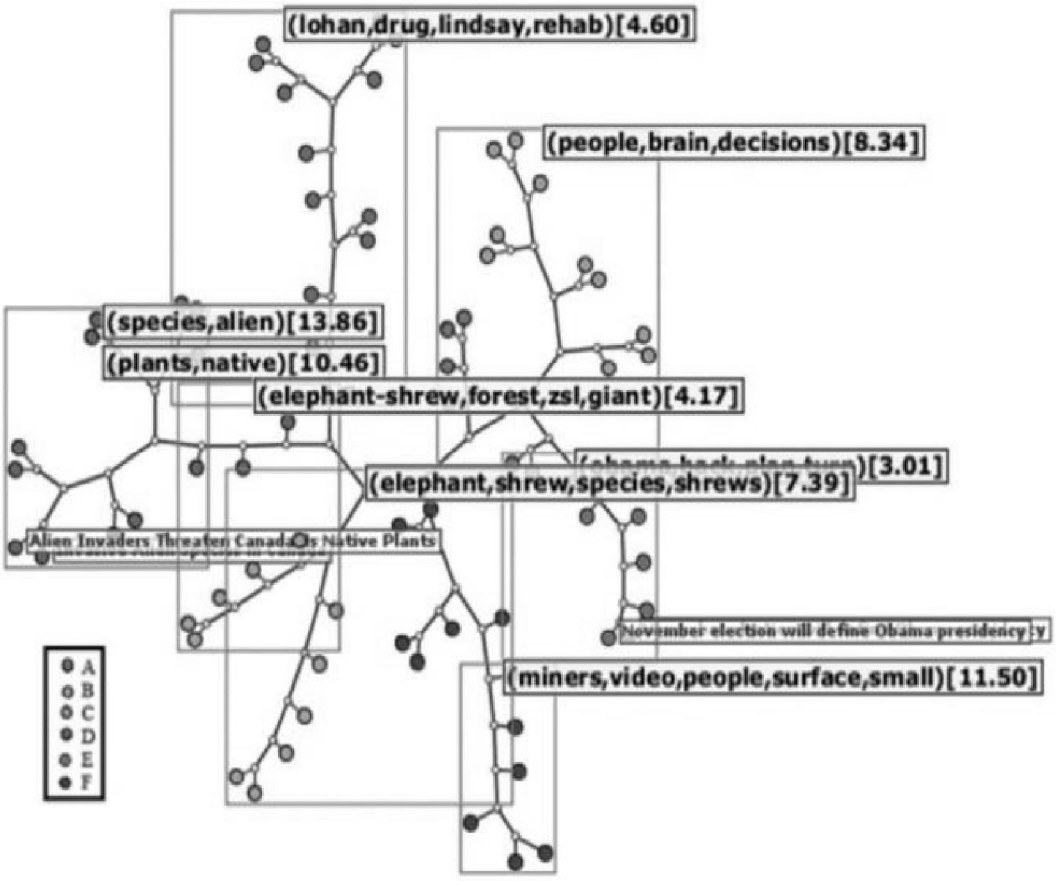

Figura 3

Mapa de textos del conjunto NEWS. Se puede ver que la técnica NJ-T es capaz de agrupar textos basados en contenido en regiones bien definidas. Los resultados son respaldados con la técnica de agrupamiento K-Means.

Por todo lo expuesto anteriormente queda demostrado que el árbol es capaz de separar noticias de contenido similar en subárboles, para lo cual se colocan los nodos con asuntos similares en la misma rama o en ramas próximas; los temas específicos tratados dentro de un tema general son divididos en subramas.

La figura 4 muestra la construcción de un mapa de imágenes NJ-T para el conjunto NEWS; el mapa emplea como marcas visuales las imágenes de noticias en miniatura. Se puede ver que las imágenes del mismo tipo son apropiadamente ubicadas en la misma rama o ramas cercanas (por ejemplo, las imágenes de Lindsay Lohan), lo cual demuestra que las imágenes similares pertenecen a la misma rama, y que las que poseen un alto grado de similitud quedan ubicadas en sus extremos.

Figura 4
Mapa de imágenes del conjunto NEWS que emplea la técnica NJ-T y se sirve de las imágenes representativas de cada noticia como marcadores visuales.

Dado que nuestro objetivo es fusionar las imágenes con los textos como si se tratara de un solo documento, se fusionaron las imágenes con los vectores de características de cada texto. El resultado aplicado al conjunto de noticas se puede ver en la figura 5. A pesar de que las características de las imágenes no superan las textuales, como se puede apreciar en las figuras anteriores (figuras 3 y 4), podemos utilizar la combinación de ambas para mejorar los resultados cuando hacemos uso de documentos multimedia.

Analizando los resultados obtenidos podemos decir que los documentos en los extremos de las ramas gozan de una similitud alta. Podemos apreciar en el mapa de imágenes que incluso existen fotos similares o iguales que fueron referenciadas en las mismas noticias, y cuyo contenido textual es parecido pero no igual. Esto ocurrió, por ejemplo, con las noticias sobre las elecciones legislativas en los Estados Unidos, donde el tema central es el respaldo del congreso al presidente Barack Obama. En la figura 5 se aprecian en la rama marcada con la letra *A* dos noticias muy similares que tratan acerca del desempeño del presidente Obama y de las expectativas que tiene con estas elecciones, donde ambas hacen referencia a fotos muy parecidas. Las dos noticias aparecen posicionadas en los nodos hojas de la misma rama, lo que quiere decir que poseen un alto grado de similitud.

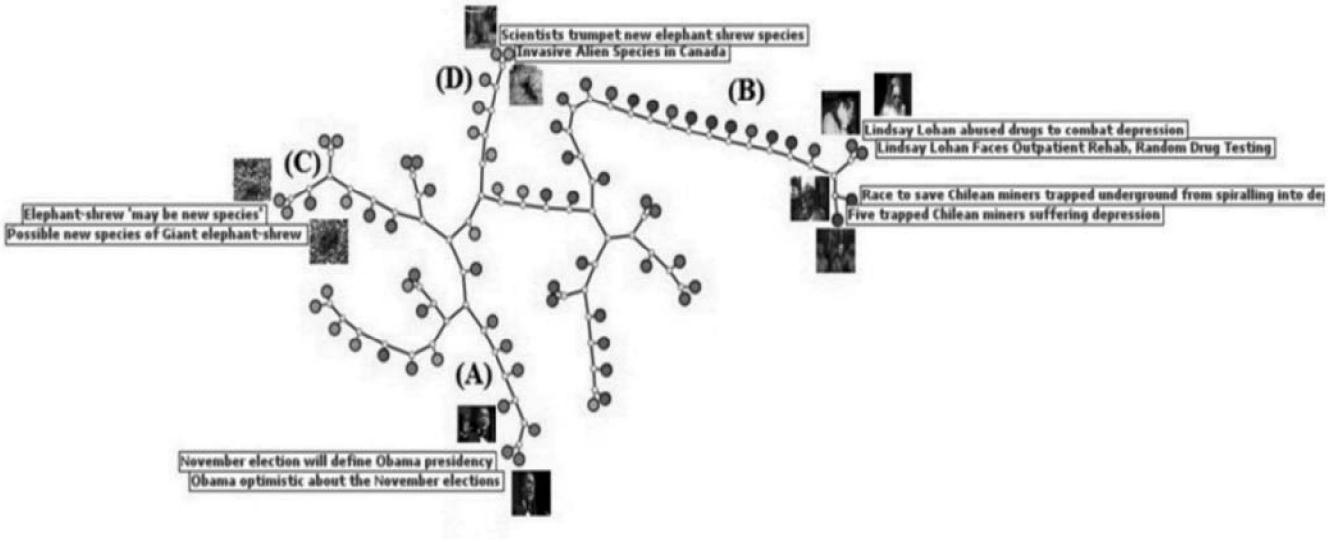

Figura 5

Visualización y exploración de la proyección realizada sobre el conjunto de noticias NEWS fusionando el conjunto de características de las imágenes y el texto; como medida de similitud se emplea la distancia del coseno.

En el extremo de la rama de la región resaltada con la letra *B* observamos que las noticias sobre el rescate de los mineros en Chile y el arresto de Lindsay Lohan parecen estar mal posicionadas. Sin embargo, tras analizar el contenido se puede apreciar que tienen en común tratar sobre la depresión, y que debido a ello guardan relación: en el caso de los mineros las dos noticias tratan sobre la depresión que podrían sufrir por estar bajo tierra varios meses; en el caso de Lindsay Lohan tratan el tema de que muchas personas famosas son víctimas de la depresión, lo que a veces las hace caer en las drogas; así mismo, se habla de cómo es su proceso de rehabilitación. Si bien en el mapa textual de

noticias presentado en la figura 3 los nodos analizados están en ramas vecinas, y los nodos están ubicados en el final de las ramas, creo que el nuevo posicionamiento está marcado por las características extraídas de las imágenes. Además, se puede observar que la rama *B* no presenta ramificaciones porque no existen temas específicos dentro de las noticias tratadas.

Las noticias de los nodos extremos de la rama *C* tratan de una especie animal descubierta en el África: los macroscelídeos o *musarañas elefante*. Esta nueva especie pertenece a la clase de los mamíferos placentarios, y habita exclusivamente en el mencionado continente. La mayoría de noticias presenta alguna de las fotos tomadas por los propios científicos. En la rama *D* los nodos ubicados en el extremo pertenencen a clases diferentes. La noticia del nodo izquierdo (amarillo) trata de la nueva especie de elefante, cuya imagen fue capturada en su hábitat natural, mientras que la noticia del nodo derecho (verde) aborda el peligro en que se encuentran las especies oriundas de Europa y América, así como con sus respectivos ecosistemas, al ser invadidas por especies exóticas foráneas.

Por todo lo expuesto se puede concluir que la técnica NJ-T aplicada a la fusión de documentos multimedia consigue posicionar documentos con un alto grado de similitud en los extremos de las ramas.

Estos resultados indican que los mapas visuales de documentos multimedia multidimensionales se pueden analizar en un único espacio de visualización, y que se pueden fusionar los vectores de características de imágenes y textos en uno solo para documentos que poseen un alto grado de similitud. El método propuesto permite posicionar los nodos de tal manera que los similares aparezcan colocados en las mismas ramas o en ramas contiguas, mientras que los nodos disímiles aparezcan distantes, próximos al centro del árbol.

Es importante resaltar que los resultados de la construcción de mapas interactivos usando la técnica NJ-T en un espacio visual están limitados por el cálculo de la matriz de similitudes. Si las distancias en dicha matriz son capaces de generar buenas distinciones entre los documentos multimedia por contenido, la técnica NJ-T es capaz de relacionar apropiadamente documentos similares. Por ello es necesario seguir explorando la adecuada representación del vector de fusión de características de los datos multimedia multidimensionales.

La contribución más importante de este trabajo con respecto a los otros es que permite integrar datos heterogéneos (documentos textuales, imágenes, y documentos con texto

e imágenes) en un único espacio visual de análisis, lo cual permite al usuario explorar los resultados, interactuar y navegar en ellos.

Referencias

Anderberg, M. R. (1973). *Cluster analysis for applications*. Nueva York: Academic Press.

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Drischl, G., Kappe, F., . . . Tochtermann, K. (2002). The InfoSky visual explorer: exploiting hierarchical structre and document similarities. *Information Visualization, 1*(3-4), 166-181. doi: 10.1057/palgrave.ivs.9500023

Ankerst, M., Keim, D., y Kriegel, H. (1996). Circle segments: a technique for visually exploring large multidimensional data sets. *Proceedings of the IEEE Visualization VIS '96*.

Bachmaier, C., Brandes, I., y Schlieper, B. (2005). Drawing phylogenetic trees. En X. Deng y D. Du (Eds.), *Proc. Intl. Symp. on Alg. and Comp. ISAAC 2005* (Vol. 3827, 1110-1121).

Becks, A., Seeling, C., y Minkenberg, R. (2005). Benefits of document maps for text accesss in knowledge management: a comparative study. *Proceedings of the 2002 ACM symposium on Applied Computing (SAC '02)*. Nueva York.

Card, S., Mackinlay, J., y Shneiderman, B. (1999). *Reading in information visualization: using vision to think*. San Francisco (California): Morgan Kauffmann.

Cuadros, A., Paulovich, F., Minghim, R., y Telles, G. (2007). *Point placement by phylogenetic trees and its application for visual analysis of document collections*. Ponencia presentada en el IEEE Symposium Visual Analytics Science and Technology (VAST 2007), Sacramento (California).

Chen, C. (2006). *Information visualization: beyond the horizon*. Secaucus (Nueva Jersey): Springer-Verlag.

Eler, D., Nakazaki, M., Paulovich, F., Santos, D., Andery, G., Oliveira, C., . . . Minghim, R. (2009). *Visual analysis of image collections*. Nueva York: Springer-Verlag.

Kalva, P., Enembreck, F., y Koerich, A. (2007). Web image classification based on the fusion of image and text classifiers. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 1, 561-568).

Luhn, H. (1968). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

MacQueen, J. B. (1967). Some methods for classification and alalysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (281-297).

Magalhães, J., e Iria, J. (2009). Exploiting cross-media correlations in the categorization of multimedia web documents. *Proceedings of the Workshop on Cross-Media Information Access and Mining at the 21st International Joint Conference on Artificial Intelligence*. Pasadena.

Manjutah, B., Ohm, J., Vasudevan, V., y Yamada, A. (2001). Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology, 11*(1).

Paulovich, F., y Minghim, R. (2006). Text map explorer: a tool to create and explore document maps. *Proceedings of the 10th International Conference on Information Visualization (IV '06)*. Washington D. C.: IEEE Computer Society.

Paulovich, F., Nonato, L., Minghim, R., y Levkowitz, H. (2008). Least square projection: a fast high precision multidimensional projection techinque and its application to document mapping. *IEEE Trans. Vis. Comput. Graph., 14*(3).

Paulovich, F., Oliveira, M., y Minghim, R. (2007). The projection explorer: a flexible tool for projection-based multidimensional visualization. *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*. Washington D. C.

Petrushin, V., y Khan, L. (2007). Multimedia data mining and knowledge discovery. En V. Petrushin y L. Khan (Eds.), *Multimedia data mining and knowledge discovery*. Londres: Springer.

PNNL. (2010). IN-SPIRE™: Pacific Northwest National Laboratory. Disponible en: http://www.pnnl.gov/infoviz

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Saitou, N., y Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425.

Telles, G., R, R. M., y Paulovich, F. (2005). *Visual mapping of text collections using and approximation of Kolmogorov complexity*. Informe técnico. ICMC-USP. São Carlos (São Paulo)

Vapnik, V. (1995). *The nature of statistical learning theory*. Nueva York: Springer-Verlag.

Wise, J. (1999). The ecological approach to text visualization. *Journal of the American Society for Informatioin Science*.