Visualizando una Red de Correo Electrónico

Lidia Marina López
Departamento de Ciencias de la Computación
Universidad Nacional del Comahue
llopez@uncoma.edu.ar

Resumen

El principal objetivo de la visualización multidimensional de datos multivaluados es la representación gráfica adecuada tanto de los datos con parámetros múltiples como de las tendencias y las relaciones clave que existen entre ellos. Diferentes características y propiedades de los datos pueden cambiar la manera en que se presente la visualización, pero no sus objetivos. El área de redes brinda una aplicación interesante y compleja y es una excelente posibilidad de cubrir la distancia entre las propuestas teóricas y las aplicaciones de esas técnicas para establecer así el valor de las mismas. En este trabajo se presenta la aplicación de técnicas de visualización de datos multidimensionales basadas en ejes no ortogonales para la representación de redes y se utiliza como caso de estudio una Red de Correo Electrónico mostrando cómo las mismas permiten una visualización efectiva. Para ello se utiliza una técnica de mapeado de datos crudos a formas visuales sobre un conjunto grande de datos. En particular se apunta al análisis y a la representación de la información generada por los mensajes de correo electrónico en un contexto determinado.

Palabras Clave: visualización multidimensional de datos, estructuras visuales, coordenadas paralelas, redes de comunicación, red de correo electrónico

1 INTRODUCCIÓN

En la actualidad, el intercambio de datos a través de medios electrónicos crece rápidamente generando grandes redes ricas en información. En paralelo con el incremento del tamaño de las redes, se han ido desarrollando métricas para el análisis de estos datos. El software tradicional de análisis de red no puede manejar el tamaño actual de las redes y todas sus capacidades de colección de datos. Así, el desafío de la visualización de este tipo de datos es crear metáforas para representar la información producida por las redes de comunicación, que permitan de manera natural e intuitiva brindar información acerca de esos datos.

Estas redes contemplan, como caso particular, la red de correo electrónico. Estos datos están implicados directamente con las organizaciones. La gran cantidad de datos producidos por este tipo de red hace que su visualización permita establecer patrones de comunicación importantes para la eficiencia organizacional.

En [1] los autores presentan la visualización de la red de correo electrónico del departamento de la empresa a la que pertenecen.

En la figura 1 se puede ver el agrupamiento en el centro, alrededor de dos fuentes importantes de comunicación. Si bien la representación es atractiva, no permite establecer patrones de comunicación que describan el comportamiento de la organización estudiada aunque sí permite clasificar el tipo de tareas que se realizan -por el color de los nodos- y el volumen de mensajes intercambiado

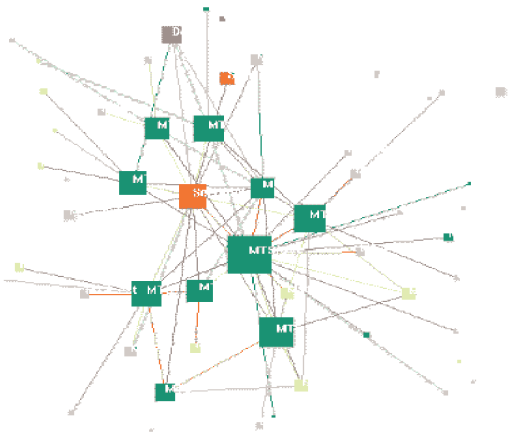

Figura 1: Red de correo electrónico - Fuente: Eick and Wills AT&T Bell Lab. [1]

-por el color de los enlaces-.

2 MAPEADO

Los Datos Crudos representan la información sin ningún tipo de tratamiento y potencialmente visualizable, es decir, representan la información del mundo real. Estos datos son formalizados en relaciones o conjuntos de relaciones que son estructurados permitiendo mapearlos a formas visuales apropiadas. El modelo de referencia utilizado para procesar datos crudos es su transformación en Tablas de Datos que mapean a una Estructura Visual que genera la visualización propiamente dicha.[3]

Las Tablas de Datos están basadas en una organización tabular determinada; las Estructuras Visuales están basadas en propiedades gráficas efectivamente procesadas por la visión humana. Si bien el dato crudo puede ser visualizado directamente, las Tablas de Datos son un importante paso intermedio cuando la información es mucha y/o abstracta, sin una componente espacial directa.

2.1 Datos Crudos

En los mensajes electrónicos existen datos fáciles de identificar en los encabezados por medio de líneas como *From:, To:, Subject:* y otros datos cuya clasificación es más compleja.

Dada la siguiente información tabulada correspondiente a 500 mensajes de un servidor de correo electrónico perteneciente a una institución educativa, se aplica la metodología de mapeado de datos a alguna forma visual para intentar obtener algún perfil de comportamiento.

Mensajes	M1	M2	M3	...
Origen	unpata	neunet	unpata	...
Destino	uncoma	infovia	yahoo	...
Tamaño	3	18	24	...
Tema	Artic..	Res..	Re:ah..	...
Contenido	*texto*	*texto*	*texto*	...

Tabla 1: Datos Crudos de mensajes electrónicos

La tabla 1 presenta los atributos dominio de origen, dominio de destino, tamaño, tema y contenido.

2.2 Transformaciones de Datos Crudos a Tabla de Datos

Se observa que existen pocos datos cuantitativos y que los atributos *Origen* y *Destino* pertenecen a un dominio común.

Como la naturaleza de la red que se está analizando es educativa, se fijan criterios para clasificar y cuantificar los datos cualitativos. Los atributos *Tema* y *Contenido* deben tener un tratamiento particular donde se establezca claramente la forma de agregación. Se adopta el criterio de contar con dos diccionarios de datos con palabras de longitud mayor que tres, que determinan temáticas. Se toma un diccionario con términos administrativos y otro con términos académicos. Los atributos *Tema* y *Contenido* se preprocesan contabilizando solamente las palabras de longitud mayor que tres resultando en tres nuevos atributos cuantitativos.

Mensajes	M1	M2	M3	...
Origen	unpata	neunet	unpata	...
Destino	uncoma	infovia	yahoo	...
Tamaño	3	18	24	...
Pal Adm	5	1	8	...
Pal Acad	2	5	4	...
Otras Pal	5	16	3	...

Tabla 2: Discretización de datos cualitativos

De esta manera se pueden obtener transformaciones de los Datos Crudos que permitan asignar alguna Estructura Visual adecuada. Debe definirse qué información sería útil para generar una representación que permita reconocer comportamientos y relaciones.

3 FORMAS VISUALES

En la Sección 1 se describe una representación de los datos correspondientes a mensajes de correo electrónico. Para el presente trabajo, si bien se tiene en cuenta la cantidad de mensajes enviados y recibidos, se ha establecido inicialmente que se busca encontrar un patrón para analizar el comportamiento de una organización a través de las características cualitativas de los mensajes que se intercambian entre las personas que la componen, no siendo necesario, en un principio, la utilización de alguna metodología de posicionamiento de los nodos y/o enlaces. Se elige utilizar una Estructura Visual diseñada para modelar relaciones.

3.1 Coordenadas Paralelas

La técnica de Coordenadas Paralelas [2] ha sido creada con el objetivo de visualizar problemas multidimensionales/multivaluados sin pérdida de información. Esta técnica es inherentemente n-dimensional dando al usuario la posibilidad de observar la relación entre n atributos. En principio, éstos pueden mapearse a n ejes y puede utilizarse el color para mapear información adicional; también podrían mapearse $n-1$ atributos a $n-1$ ejes y el atributo adicional a color.

La ventaja de los ejes paralelos sobre los ortogonales es el hecho de que sus limitaciones están basadas en el tamaño del área de representación disponible. Su utilización no produce pérdida de información. Cada observación en un conjunto de datos está representada como una serie de segmentos de línea que intersectan los ejes verticales, cada uno de los cuales escala a un atributo diferente. Los valores del atributo para cada observación es marcado en cada eje relativo entre de los valores máximos y mínimos del atributo para todas las observaciones; los puntos son conectados usando segmentos de línea. El resultado es una sucesión de segmentos de línea a través de n dimensiones para cada observación.

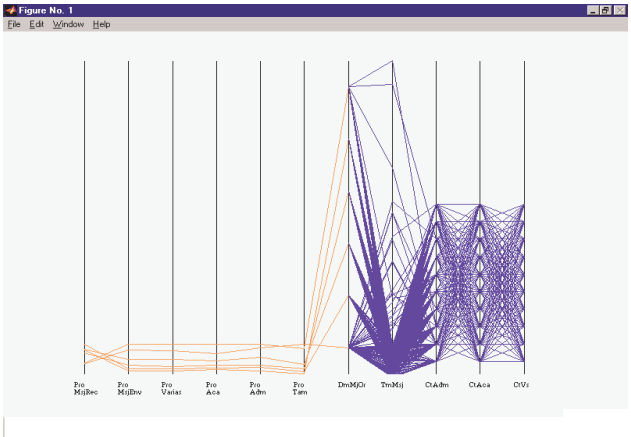

Figura 2: 500 Mensajes en Coordenadas Paralelas

3.2 Análisis de la representación

En la Figura 2 están representadas once dimensiones. Las primeras seis corresponden a valores derivados de los mensajes agrupados por dominio, de color naranja. Las últimas cuatro corresponden a los valores de los mensajes, de color azul. En el *eje* 7 están ubicados los dominios. Las dimensiones representan los siguientes atributos, de izquierda a derecha:

Valores agrupados por Dominio de Origen:

- 1. Cantidad total de Mensajes Recibidos.
- 2. Cantidad total de Mensajes Enviados.
- 3. Porcentaje de Palabras Varias. De todas las palabras contenidas en el título y el contenido del mensaje, se calcula el porcentaje de las que pertenecen a la lista de palabras elegidas como administrativas sobre el total de palabras que han sido clasificadas por alguno de los tres tipos.
- 4. Porcentaje de Palabras Académicas. Idem anterior.
- 5. Porcentaje de Palabras Administrativas. Idem anterior.
- 6. Porcentaje de Tamaño. Con respecto al total de Mensajes para cada dominio de origen.

7. Dominio Origen.

Valores de cada Mensaje:

- 8. Tamaño del mensaje.
- 9. Cantidad de Palabras Administrativas.
- 10. Cantidad de Palabras Académicas.
- 11. Cantidad de Palabras Varias.

La Figura 2 predispone a un escepticismo saludable, ya que la gran cantidad de líneas provoca desaliento. Para poder comenzar a analizar la Estructura Visual, es imprescindible contar con la interacción adecuada. En los ejes del 1 al 6 se representan los valores de los mensajes agrupados por el dominio que origina el mensaje. En los ejes del 8 al 11 se representan los atributos de cada mensaje: tamaño, cantidad de palabras administrativas, cantidad de palabras académicas y cantidad de palabras varias.

Teniendo en cuenta los rangos de valores de cada eje y el objetivo de la representación, se trata de ver si existen patrones que brinden pistas para encontrar relaciones. En este caso se busca encontrar qué tipo de mensajes se envían dentro de la organización elegida. El resultado ideal sería que una cantidad destacada de mensajes se clasifiquen dentro del tipo académico, es decir, mayor cantidad de mensajes con palabras académicas y que al menos el dominio de origen correspondiente a la Universidad, *uncoma*, tenga esta característica.

Los valores que corresponden al tipo académico están representados en los ejes 4 y 10. Por lo tanto debe concentrarse la atención en estos ejes siendo útil discriminar la representación por dominio de origen.

3.3 Coordenadas Paralelas Mejoradas

Como los Dominios de Origen representan un atributo independiente, se propone buscar una forma de representarlos fuera del diagrama de ejes. Se busca representar los dominios como íconos de forma rectangular. Es necesario redefinir la Tabla de Datos donde se establecen la información a ser representada por el ícono y la información de los mensajes para las coordenadas paralelas.

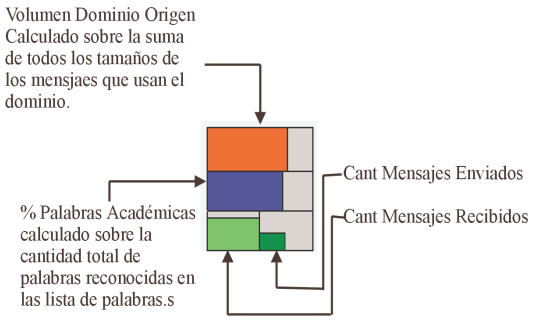

Figura 3: Icono para representar el Dominio de Origen

La Figura 3 muestra la forma básica del ícono y el significado de las partes.

Para los mensajes se utilizan las coordenadas paralelas. Es claro que una interacción que debe permitirse es el intercambio de la posición de los ejes buscando evitar que las líneas se aglomeren ya que esto dificulta la visualización.

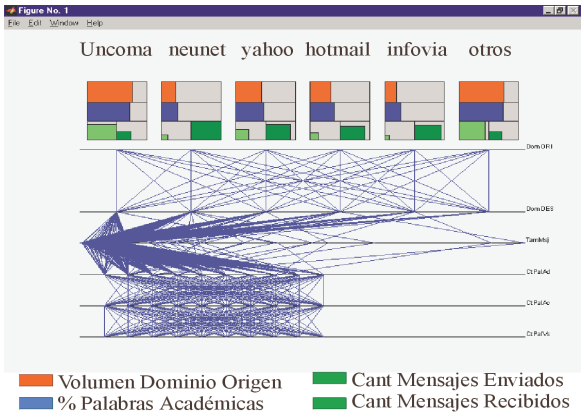

Figura 4: Implementación Coordenadas Paralelas Con detalle de nodos - Matlab

La Figura 4 muestra la Estructura Visual utilizada. Una interacción adecuada en esta etapa es la posibilidad de seleccionar un punto en los ejes y resaltar todas las líneas que desde allí nacen.

4 CONCLUSION

El prototipo presentado brinda una alternativa para la visualización de datos con estructura de red donde la Tabla de Datos contiene información de los nodos de la red por un lado e información de los enlaces por otro. En el caso de estudio, la Tabla de Datos está compuesta por dos estructuras: *Dominios* y *Mensajes*.

Se concluye que, a partir del modelo de referencia aplicado, el cual implica el mapeado de los datos a una estructura visual, se obtiene una visualización adecuada de la red de correo electrónico que permite una evaluación coherente del comportamiento de la organización estudiada.

5 TRABAJO FUTURO

El trabajo futuro apunta a la visualización de información de redes sociales, desarrollando inicialmente tareas se detallan a continuación.

-En cuanto a los datos a explorar, se está trabajando con redes de datos sociales que tienen alguna topología, buscando la mejor manera de desarrollar la Tabla de Datos para luego aplicar el prototipo desarrollado.

-En cuanto al modelo de referencia, se intentará trabajar en producir Tablas de Datos que brinden mayor información que permita decidir la Estructura Visual a utilizar.

-En cuanto a la visualización propiamente dicha, se seguirá trabajando sobre técnicas orientadas a ejes no cartesianos y al agregado de herramientas para análisis interactivo.

Referencias

[1] Stephen G. Eick and Graham J. Wills. Navigating large networks with hierarchies. In *Proc. IEEE Conf. Visualization*, pages 204–210, 1993.

[2] Alfred Inselberg. Multidimensional detective. pages 100–107. IEEE Computer Society, 1997.

[3] B.Shneiderman S.Card, J.Mackinlay. *Readings in Information Visualization - Using Vision to Think*. Morgan Kaufmann Publisher Inc 1st edition, San Francisco, California, 1999.