"Visualización de Documentos: Un ambiente para la visualización de noticias"

Aizemberg A., Casullo P., Mislej E., Paoletta M., Pardieux E., Picardi G., Santos M. y Tylim H.
{aa7g, pcasullo, emislej, mpaolett, epardieu, gpicardi, msantos, hptylim}@dc.uba.ar

Departamento de Computación - FCEyN - UBA
Planta Baja - Pabellón I - Ciudad Universitaria
(1428) Ciudad Autónoma de Buenos Aires. Argentina.
Tel.: +54-11-4576-3390/96 int 701/702. Tel/Fax: +54-11-4576-3359.

Palabras claves: *visualización de información, visualización de documentos, visualización de noticias.*

Resumen

La disciplina de visualización de información ha ganado la atención de muchos expertos e investigadores en los últimos años. Esta disciplina comprende la exploración de vastos y complejos espacios de información, por lo tanto, proveer de ayudas de orientación intuitivas y técnicas de navegación es esencial para permitir la exploración y el reconocimiento de tendencias y relaciones ocultas en los datos.

En este trabajo presentamos un ambiente de visualización para noticias de los medios gráficos. El objetivo que perseguimos durante el desarrollo fue la creación de un ambiente amigable y potente. Amigable en el sentido cuya operación sea intuitiva y no demande al usuario más que un mínimo de aprendizaje. Potente en el sentido de ofrecerle al usuario la posibilidad de explorar las noticias por medio de distintas visualizaciones, sin que por ello se pierda la coherencia entre un salto de vista a otro.

La herramienta permitirá explorar las noticias por fecha, sección y por temática, así como también brindar la posibilidad de interacción con el fin de lograr una mayor comprensión y capacidad de análisis del conjunto seleccionado de datos.

1 Introducción

Existen varios y diversos diarios en la web [1, 2, 3] que en la actualidad brindan la misma o casi la misma información que en su formato impreso. En ambos casos es difícil clasificar las noticias de otra forma que no sea la predeterminada por las distintas ediciones, ya sea para hacer una lectura comparativa, paralela, o inclusive una mínima investigación sobre algún tema en particular.

Asimismo, para responder a preguntas del estilo: durante cuánto tiempo se habló de tal tema o cómo evolucionó tal tema en el transcurso de los días; las personas recurren a algún programa periodístico por televisión o, quienes son capaces, leen los diarios entre líneas. Cada quien saca sus propias conclusiones de manera subjetiva, y ocurre, que por lo general, se tiene poca buena memoria sobre los eventos ocurridos y esto repercute en dichas conclusiones.

Lo que se propone en el presente trabajo es utilizar los titulares de los diarios para construir una serie de visualizaciones que permitan responder de manera objetiva a las preguntas anteriores de manera rápida e intuitiva.

2 Desarrollo

La herramienta es esencialmente una aplicación cliente-servidor, la cual se encuentra en fase de desarrollo. El cliente utiliza Scalable Vector Graphics (SVG) [4] (un lenguaje para describir gráficos en dos dimensiones basado en XML) embebido en HTML y funciones javascript para las interacciones.

El siguiente ejemplo muestra un círculo pintado de rojo con borde azul especificado en SVG.

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 20010904//EN"
    "http://www.w3.org/TR/2001/REC-SVG-20010904/DTD/svg10.dtd">
<svg width="12cm" height="4cm" viewBox="0 0 1200 400"
     xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink">
<circle cx="600" cy="200" r="100" fill="red" stroke="blue" stroke-width="10" />
</svg>
```

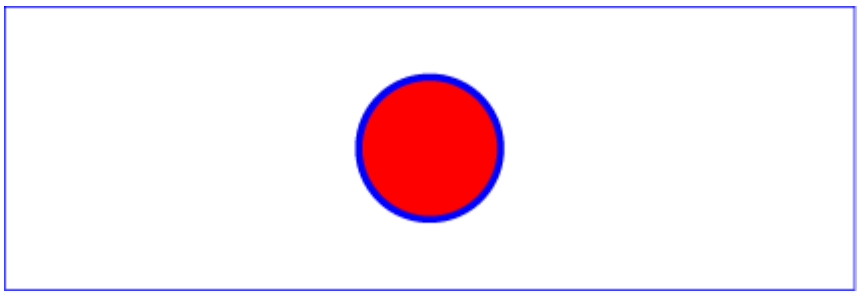

Figura 1

El servidor genera dinámicamente los archivos SVG, con los parámetros de las consultas definidos por el usuario, utilizando tecnología JSP, y extrayendo la información de una base de datos Mysql.

Conjuntamente con la imagen, se envían funciones javascript que son las encargadas de efectuar las interacciones. Gracias a la tecnología SVG, es posible contar con interacciones más sofisticadas con un tiempo de respuesta óptimo, ya que la lógica de presentación se encuentra en el cliente.

En resumen, con la arquitectura propuesta, una visualización se reduce tan sólo a uno o varios archivos JSP, que generan archivos SVG en forma dinámica. Por lo tanto, se podrán agregar tantas visualizaciones como se quiera, y con un mínimo trabajo de integración en la aplicación.

Para poder utilizar esta aplicación es necesario instalar previamente algún visor de SVG, el *plug-in* de Adobe [5] actualmente es el más popular, tener una conexión a Internet y utilizar el modo de video en 1024 x 768.

3 Descripción del ambiente

El ambiente de trabajo desarrollado admite la posibilidad de contener varias visualizaciones. En esta primera se han desarrollado dos: *ThemeRiver* y *ReadingTimeNews*.

El espacio de representación está ocupado por los siguientes componentes:

- **Vista principal**: en este marco se realizan las principales interacciones, además de contener al gráfico principal.
- **Parámetros generales**: lugar donde se especifican variables comunes a todas las visualizaciones y otras operaciones de selección y creación de nuevas palabras clave.
- **Zoom in, Zoom out & Panning**: desde aquí se puede tener una visión de toda la información seleccionada y moverse libremente dentro del contexto general.
- **Información visual de 1er nivel**: al seleccionar una zona en la vista principal se presenta mayor nivel de detalle permitiendo interacciones más especificas.
- **Información textual de detalle de 1er nivel**: aquí se realiza un tipo de zoom semántico.
- **Información textual de detalle por demanda (2do nivel)**: este último recurso es el más parecido a la lectura normal de la nota.
- **Leyenda y barra de estado**: elementos característicos en cualquier gráfico o aplicación.

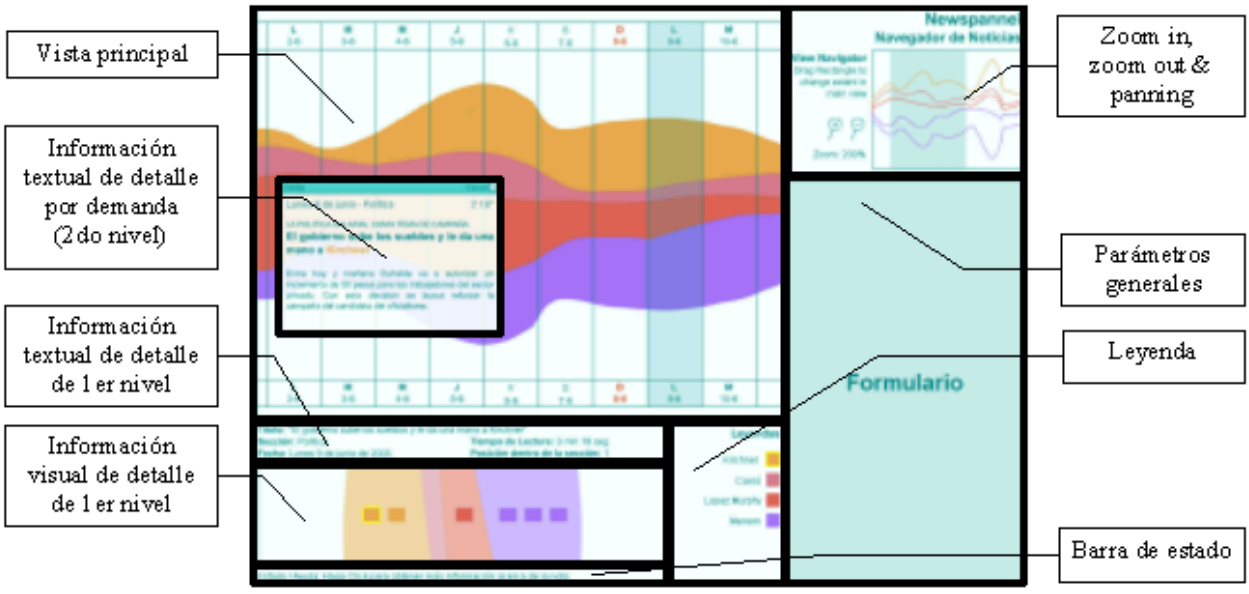

Figura 2

4 Especificación

Para especificar las visualizaciones desarrolladas lo haremos de acuerdo a la taxonomía propuesta por Chi [6]. Cabe destacar que todas las visualizaciones comparten los datos crudos *(raw data)* y el modelo de datos subyacente.

- ***Datos crudos:***

La información proviene de un conjunto de páginas web (html), cada una correspondiente a una edición del diario. A través de un robot de búsqueda, las páginas son descargadas de su sitio y transformadas a uno o varios archivos de texto plano delimitado por comas.

- ***Modelo de datos:***

Este modelo de datos se parece a un modelo estrella [7] comúnmente utilizado en los repositorios para hacer análisis multidimensional. En nuestro caso la tabla de hechos son las noticias y las dimensiones son las fechas, las secciones y las temáticas. Con este modelo de datos es sencillo agregar jerarquías y tablas de sumarización permitiendo que las visualizaciones no recorran todos los datos cuando se busca información sumarizada.

El dominio de las secciones son las siguientes: Política, Economía, Opinión, Internacionales, Sociedad, Deportes, Espectáculos y Extras.

La definición de una temática va a estar dada por el usuario a través de una serie de palabras claves. Por ejemplo el tema "La guerra de Iraq", podría estar definido por las palabras clave Iraq, Guerra, Bush y Hussein. Entonces si existe una noticia que contenga alguna de esas palabras clave, la herramienta interpretará que esa noticia habla de "La guerra de Iraq". El usuario podrá definir tantas temáticas como desee.

- ***ThemeRiver; abstracción visual, vista e interacciones***

Se ha implementado una versión libre del *ThemeRiver*^{TM} [8], que muestra el río simétrico con las corrientes en diferentes colores según temática o sección y agrega totales como el tiempo de lectura de cada corriente, la cantidad de artículos o titulares por sección o por temática, el orden del artículo desde la tapa del diario y la posibilidad de seleccionar un rango de fechas que afecta a todas las corrientes definidas en el río.

Las noticias no se representan directamente a ninguna estructura visual, pero sí sus atributos definidos en la abstracción de datos. En este caso, el tiempo total de lectura de cada sección o temática, se grafica en el ancho de cada corriente y los atributos de texto (título, volanta y copete), no sufren ninguna modificación.

Las agregaciones sección o temática se pintan con el color de la corriente asociada. Cabe destacar que se puede elegir un mapeo u otro, no los dos a la vez. Entonces cada corriente contendrá los artículos de acuerdo a la selección del eje vertical (sección o temática), y su ancho estará dado por la cantidad de artículos por unidad de tiempo.

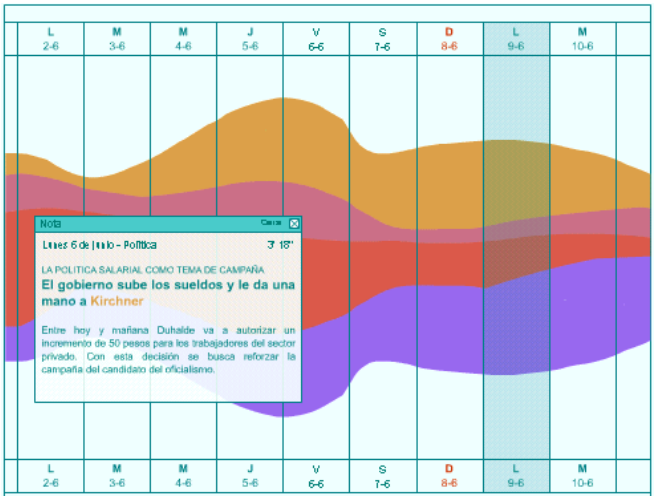

Figura 3

Una interacción posible, que no sólo afecta a esta visualización sino a todas, es la de requerir información textual por demanda para llegar finalmente a la lectura del documento. Se puede acceder a dicha noticia en una ventana flotante o bien a través de la web recurriendo a las ediciones electrónicas de los distintos medios gráficos.

- ***ReadingTimeNews; abstracción visual, vista e interacciones***

La idea central de esta visualización, es poder apreciar las sucesivas ediciones de un diario para un rango de fecha dado, en forma parcial o total, agrupadas por sección. Se utiliza al eje *y* como eje temporal.

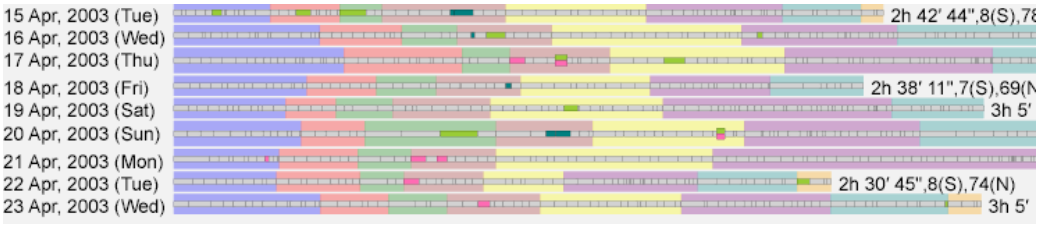

Figura 4

El nombre de esta visualización fue bautizado como *ReadingTimeNews*, debido a que la proporción de las noticias está dada por el tiempo de lectura de las mismas, se puede apreciar a la derecha de la figura 4 el tiempo total de lectura de cada edición.

En la misma figura se pueden observar con distintos colores a las 8 secciones y también destacados con otro color a las notas que contienen alguna palabra clave buscada.

Cada sección se encuentra fragmentada en las noticias que la componen, donde el orden de una noticia, es el orden original de aparición en dicha sección. Dado que en esta visualización nos interesa comparar secciones, proponemos dos elementos visuales para dicho fin. La idea es poder fijar a dos secciones de interés en estos dos elementos y poder apreciarlas con mayor detalle, realizando un zoom semántico. También es posible fijar sólo una sección de interés y compararla con las secciones restantes, simplemente señalando con el cursor a dichas secciones.

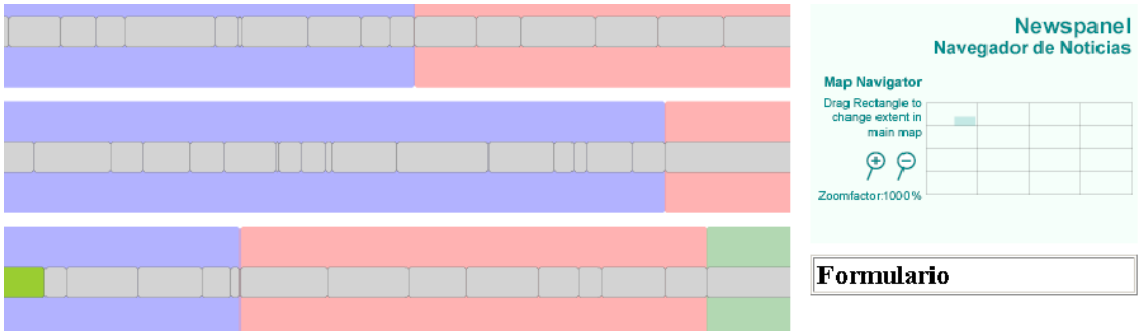

Figura 5

En la figura 5 se puede observar la vista principal afectada por un factor de escala mayor. En este caso se muestran secciones que tienen noticias cuyo tiempo de lectura es casi nulo, sin embargo en esta escala, se pueden acceder fácilmente.

5 Caso de estudio: "La guerra de Iraq"

Para este caso se definió la temática "guerra de Iraq" asociando las siguientes palabras clave: guerra, Iraq, Bush y Hussein.

En las figuras 6 y 7 muestran las ediciones del diario Clarín en el rango de fechas del 2/1/2003 al 15/1/2003 y del 16/3/2003 al 31/3/2003.

A primera vista y sin pretender efectuar un análisis de política internacional, se puede observar la diferencia de magnitud de la sección Internacionales entre el primer rango de fechas y el segundo. Internacionales prácticamente duplicó su contenido durante la guerra. También se puede ver como durante ese período, como Internacionales le sacó protagonismo a las secciones de Política y Economía.

Notar como en la figura 6 apenas aparecen las palabras claves y en la figura 7 las notas están plagadas de ellas. Es evidente que las palabras claves aparecen en mayor proporción en la sección Internacionales, sin embargo se puede observar que también aparecen con fuerte porcentaje en Opinión y en Política.

Un aspecto secundario a este análisis, pero no menos importante, es que la sección Deportes permaneció inalterable frente a las noticias de la guerra de Iraq. No cambió su interés ni hubieron demasiadas notas que hablen de la guerra.

Figura 6 – Enero de 2003, antes de la guerra de Iraq

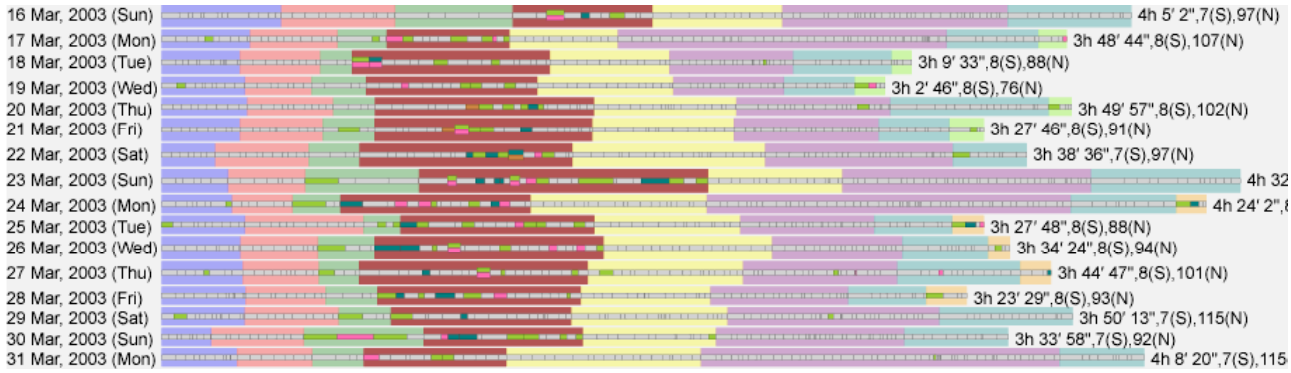

Figura 7 – Marzo de 2003, durante la querra de Iraq

Figura 8 - Referencias comunes a las figuras 6 y 7

6 Conclusiones y trabajo futuro

Las aplicaciones de visualización, en general son un escenario muy rico y apto para la incorporación de mejoras continuas. El entorno que presentamos no escapa a esta regla. Es por eso que el enfoque fue el de preparar un ambiente general para la visualización de noticias lo suficientemente flexible y versátil como para incorporar nuevas funcionalidades. Algunas de estas ideas fueron planteadas durante el diseño de la herramienta:

- **Nuevas fuentes de datos:**

Las fuentes de datos pueden extenderse a diversas publicaciones electrónicas fácilmente. Con el agregado de estas nuevas fuentes sería posible la comparación y responder a preguntas como cuánta repercusión tuvo tal tema en los distintos medios, quién fue el precursor de tal tema o quién fija la agenda periodística.

- **Text Mining:**

El dominio de noticias periodísticas es muy apto para la aplicación de técnicas de Text Mining como extracción de conceptos, clustering o agrupamiento de noticias, ranking y técnicas avanzadas de recupero de información.

- **Nuevas vistas:**

Durante el transcurso del diseño de la herramienta, muchas vistas fueron presentadas. Algunas de ellas fueron destinadas para futuras versiones. A continuación se detallan algunas de estas vistas que fueron presentadas y no llegaron a implementarse:

  - **Ramas secas:**

    En esta vista se utiliza la metáfora gráfica de ramas colgadas sobre una soga para ser secadas. Las secciones de cada día se dibujan como a líneas verticales unidas en su extremo superior. Cada línea tiene su altura proporcional al tiempo de lectura total asociada y se dibuja con el color asignado de cada sección. Las notas que contienen las palabras clave son representadas con cajitas de color, de altura proporcional a su tiempo de lectura y se colocan verticalmente según su posición relativa dentro de la sección que la contiene.

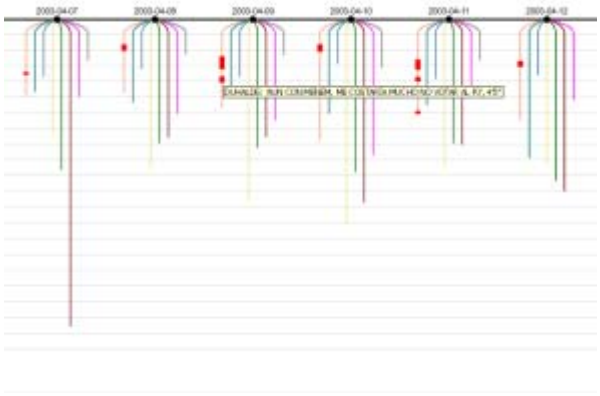

Figura 9

- ***Tree List* mejorado:**

En esta vista se utiliza la conocida estructura visual *Tree List*, que se puede ver en *Windows Explorer*. A esta se le agregan algunas mejoras visuales como la representación del tiempo de lectura y posición de las secciones y las notas que se despliegan y ocultan de manera interactiva.

- **Escritorio inteligente:**

Esta vista hace referencia a la metáfora de escritorio, donde el periodista, el analista o el simple lector acomoda las notas sobre un espacio según su criterio de relevancia. Al inicio, todas las notas se despliegan sobre el escritorio según un criterio de ordenación propuesto por la editorial.

Luego el usuario será capaz de interactuar con las notas y poder eliminarlas, cambiar su tamaño y posición. Al desplegarse gran cantidad de títulos, el usuario tendrá la posibilidad de dar un vistazo general y poder enfocarse en las noticias de su interés pudiendo con ellas darle su tratamiento particular.

Esta visualización será capaz de recordar el criterio de relevancia propuesto por el usuario y aplicarlo a nuevas ediciones. Con esto se estaría logrando la idea del periódico personalizado.

Figura 10

7 Referencias

[1] Clarín.com, http://www.clarin.com

[2] La Nación Line, http://www.lanacion.com.ar

[3] Página 12 Web, http://www.pagina12.com.ar

[4] Scalable Vector Graphics (SVG), http://www.w3.org/Graphics/SVG

[5] Adobe SVG Viewer Download Area, http://www.adobe.com/svg/viewer/install/

[6] Chi, E.H. *A Taxonomy of Visualization Techniques Using the Data State Reference Model.* Proceedings of InfoVis 2000, IEEE Computer Society, 69-75.

[7] Kimball, R. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses.* John Wiley & Sons, 1996.

[8] Havre S., Hetzler B., Nowell L., *ThemeRiver(tm): In Search of Trends, Patterns, and Relationships*. Battelle Pacific Northwest Division. Presented at IEEE Symposium on Information Visualization, InfoVis 1999.