Press-clipping

Cuaderno Red de Cátedras Telefónica

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

Nessie, un rastreador para las noticias de tu interés publicadas en prensa.

Cátedra Telefónica de la Universidad de Las Palmas de Gran Canaria

"Hay casi el doble de lectores de periódicos que internautas"

Alexis Quesada Arencibia
Jose Carlos Rodríguez Rodríguez
Javier Toledo Mediavilla
David Freire Obregón
Eliezer Talón Socorro
Víctor Álvarez Hernández
Guanchor Ojeda Hernández

Octubre 2012

www.catedras.telefonica.es
www.rcysostenibilidad.telefonica.com

Telefónica

Biografía

Alexis Quesada Arencibia

Doctor en Informática, actualmente es Profesor Contratado Doctor del Departamento de Informática y Sistemas de la Universidad de Las Palmas de Gran Canaria (ULPGC). Posee una dilatada experiencia en las áreas de docencia, investigación e innovación tanto tecnológica como educativa. Como docente ha sido profesor a tiempo completo de la ULPGC desde el año 2001 y en estos momentos es el Director del Instituto Universitario de Ciencias y Tecnologías Cibernéticas (IUCTC).

Jose Carlos Rodríguez Rodríguez

Doctorando en Informática, ejerce como Profesor en tele-formación de la Universidad de Las Palmas de Gran Canaria (ULPGC), habiendo sido profesor a tiempo parcial durante el periodo 2007-2012 de la citada universidad. Parece que una sobre-exposición pertinaz y prolongada al área del reconocimiento de caracteres (tema nuclear de su tesis) le cualificó para tomar el rol de co-tutor en la parte de Nessie que devino en el Proyecto de Fin de Carrera "Nessie, reconocedor óptico de texto en recortes de prensa escrita".

Javier Toledo Mediavilla

Ingeniero Informático especializado en Ingeniería de Software. Director técnico y fundador de TheAgileMonkeys, una empresa especializada en el desarrollo de aplicaciones distribuidas para web y dispositivos móviles. También cofundador de la startup Menu4today. Participó en el proyecto Nessie como cotutor del proyecto de fin de carrera "Nessie: Clasificador de noticias por temas".

David Freire Obregón

Ingeniero informático, actualmente es docente contratado en la estructura de teleformación de la Universidad de Las Palmas de Gran Canaria (ULPGC) así como docente de la UNED. En la actualidad posee dos másters, uno en informática y otro en procesos educativos. Es autor o coautor de un número significativo de artículos en ambas áreas.

Eliezer Talón Socorro

Ingeniero Informático nacido en Las Palmas de Gran Canaria en 1983, especializado en programación web con CakePHP y desarrollo de aplicaciones móviles para iOS. Ha participado en la fase inicial del proyecto Nessie como parte integrante de su Proyecto Fin de Carrera, desarrollando el módulo de segmentación y reconocimiento de caracteres.

Víctor Álvarez Hernández

Nació en Las Palmas de Gran Canaria en 1983, Ingeniero Informático, especializado en el desarrollo de aplicaciones web tanto en el "Back-end" con PHP y .NET como en el "Front-End" con HTML5, CSS3 y JavaScript. Dentro del proyecto Nessie participó en la primera etapa elaborando su Proyecto Fin de Carrera:"Nessie: Clasificador de noticias por temas" en el que desarrollo del módulo de clasificación temática de noticias, el controlador del proyecto global y la interfaz gráfica de usuario (incluyendo una interfaz de validación de resultados).

Guanchor Ojeda Hernández

Nació en Las Palmas de Gran Canaria en 1983. Ingeniero Informático, que desarrolla sus primeros años en la profesión como desarrollador web. Participó en la primera etapa de Nessie durante la elaboración de su Proyecto de Fin de Carrera: "Nessie, segmentación automática recurrente adaptativa". En el que se abarcó, principalmente, los dominios de la Visión por Computador y Análisis de Imágenes de Documentos.

Índice

1. Estar informados, ¿una necesidad?
2. El caso de la prensa escrita.
3. ¿Qué ofrece Nessie?
4. Controlador del proyecto Nessie.
5. Segmentación de noticias mediante análisis del *layout*.
6. Extracción de texto con técnicas OCR.
7. Clasificación de noticias por coincidencia temática.
8. Interfaz web.
9. ¿Dónde estamos y a dónde vamos?

1. Estar informados, ¿una necesidad?

¡Pues parece que sí! Nuestra realidad como seres sociales es que no vivimos aislados. En consecuencia, a casi todos nos interesa saber qué se dice de nosotros o de lo que hacemos ahí fuera, sea bueno o malo. La aprobación, el reconocimiento externo, el descrédito o la calumnia son importantes para nosotros casi desde el principio de nuestras vidas. Y si además se es, por ejemplo, político, consejero delegado de una firma comercial reconocida o artista famoso existe un interés profesional fundado: nuestro caché depende de la percepción que tengan las masas casi tanto como de nuestro propio esfuerzo.

Hoy en día los medios de comunicación juegan un papel importante no sólo por proveer un conocimiento objetivo sobre multitud de temáticas sino por influir a la opinión pública con su particular enfoque editorial. De hecho con frecuencia son criticados por ponerse al servicio de grupos empresariales o como adalides de alguna ideología política. Por otra parte son mecanismo legítimo para la denuncia de injusticias sociales o de divulgación de conocimiento científico. Lo que resulta claro es que se ha generado un gran interés en conocer de manera exhaustiva la información que en ellos aparece.

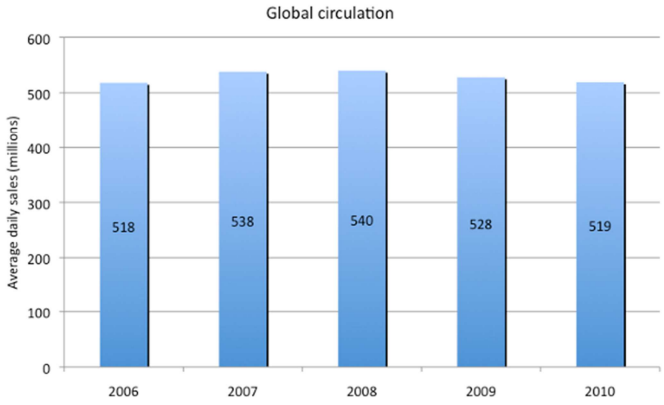
La industria del seguimiento de medios surge como respuesta a la demanda citada previamente: ofrece un servicio que recopila, filtra y clasifica la ingente cantidad de información disponible, ofreciendo un informe personalizado para cada tipo de usuario.

Pero, ¿quién querría pagar un servicio que le envía regularmente un compendio personalizado de recortes de periódico? Bien. El mercado no es pequeño. No es algo tan simple como conocer la opinión que los demás tengan de uno. La llave que abre la puerta a este modelo de negocio puede resumirse con la célebre cita atribuida a Francis Bacon: “El conocimiento es poder”.

Aceptemos como premisa que a una firma comercial le interesa tener poder en su sector para colocar su producto en una situación de ventaja con respecto a la competencia. Alcanzar ese poder supone, entre otras cosas, una demanda de conocimiento. El conocimiento se construye con información. Y la información se obtiene del análisis de datos. ¿Donde están los datos? Pues emanan de las fuentes de datos. Y aquí es donde entran en escena los periódicos porque ¡los periódicos son una fuente de datos! El problema surge porque una firma comercial, en la mayoría de los casos, no es capaz de abarcar toda la información disponible en los medios de comunicación.

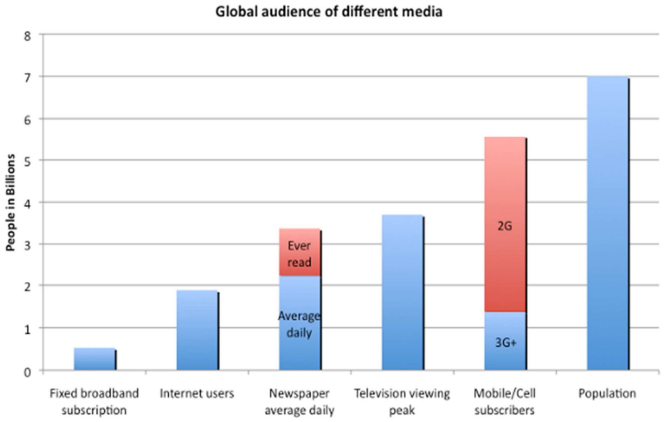
2. El caso de la prensa escrita.

Pudiera pensarse que, en los tiempos que corren, la prensa escrita como fuente de datos ha quedado obsoleta. La tendencia general a la digitalización de todo, y en concreto del contenido puramente informativo, es incuestionable. Pero, los periódicos, al igual que soportes como libros o revistas se resisten a esta transformación.



Las investigaciones llevadas a cabo por WAN-IFRA (Asociación mundial de periódicos y publicadores de noticias), reflejan que la circulación de este tipo de medio se ha visto **sólo** ligeramente reducida en los últimos 5 años.

Globally, there are nearly twice as many newspaper readers as Internet users

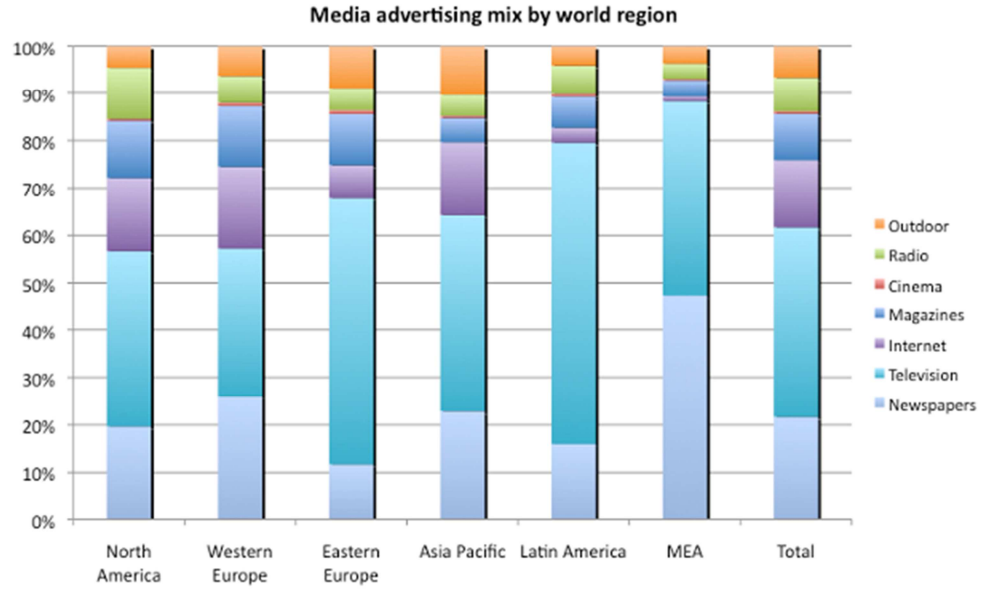


"A nivel mundial, hay casi el doble de lectores de periódicos que internautas", cita el estudio World Press Trends, realizado por la asociación anterior (la cantidad de información que pueda demandar un internauta de media, no entra como objetivo de este estudio...).

Además, esta industria sigue siendo un gran negocio ya que 10% de la inversión publicitaria es en periódicos, empatado con internet, y sólo superado sólo por el gigante de la Televisión. Esto nos permite tener una visión más pragmática a su existencia.

¿Qué queremos decir con estos datos? Pues que hay presente, y muy importante, en la divulgación de noticias en prensa escrita.

Aprovechamos para mencionar, otro de los datos arrojados por la World Press Trends: en las regiones emergentes, dónde evoluciona la educación, la búsqueda del conocimiento y las libertades de expresión se producen crecimientos importantes en la divulgación de la información por medios escritos.



Para los periódicos, no sólo existe un **presente**, sino que, con la tendencia actual, seguirá llegando a sus lectores durante muchos, muchos años. Y sólo un cambio radical en esta tendencia o/y en los demandantes de información en los próximos años (que a día de hoy no se ha dado a pesar de Google, los feeds RSS y similares, los portales de noticias, blogs, etc.), podrían alterar ese **futuro**.

3. ¿Qué ofrece Nessie?

El seguimiento de medios es un sector relativamente nuevo en Canarias, al menos en lo que se refiere a su explotación comercial por parte de empresas regionales. Pero también es cierto que está creciendo el interés por satisfacer esa necesidad real de muchas empresas de sondear los medios de comunicación y automatizar la recopilación de información de interés. Sin embargo existe una barrera económica que frena cualquier posibilidad de expansión: el alto coste de las soluciones que ya existen en el mercado.

El software que se desarrolla actualmente está pensado para grandes compañías que cuentan con clientes en todo el mundo, procesando medios de comunicación de varios países. Por ello, una empresa en la que sus clientes sólo necesitan conocer un conjunto reducido de periódicos no tiene la solvencia necesaria para implantar un sistema a gran escala. Mientras tanto, el procedimiento actual que siguen varias empresas a nivel regional es muy rudimentario: un grupo de documentalistas leen y clasifican manualmente las noticias aparecidas en diferentes periódicos, usando un soporte informático muy básico en términos de automatización.

Nessie ofrece un camino intermedio con el desarrollo de un software que permita la automatización del seguimiento de la prensa escrita. El objetivo principal es reducir el tiempo que se invierte en todo el proceso, transformando el rol actual del documentalista a un supervisor y eventual corrector de los errores que cometa el programa. Con ello pretendemos obtener un producto con una funcionalidad similar a las soluciones actuales pero a un coste reducido al omitir funciones innecesarias o poco demandadas. Así se puede proveer una solución menos compleja y costosa, asequible para empresas cuyo radio de actuación es reducido.

Como aplicación informática Nessie se compone de tres módulos interconectados que imitan las labores de un documentalista. El primer módulo identifica y recopila bloques de noticias mediante el análisis del aspecto visual y la maquetación de las páginas del periódico. El segundo módulo se encarga de obtener el texto que se encuentra en cada uno de los bloques encontrados en la fase anterior, incorporando cierta meta-información sobre distintos parámetros. Por último el tercer módulo controla el proceso de suministro de periódicos, realiza una clasificación temática de cada noticia a partir de un conjunto de palabras clave predefinido, y genera los resultados para ser entregados al usuario.

4. Controlador del proyecto Nessie.

Antes de describir los tres módulos principales de Nessie, comentados anteriormente, cabe mencionar la existencia de un Controlador de la aplicación. Este, se encargará de gestionar la interacción entre los módulos y de la cesión de los recursos necesitados por cada uno. Las tareas realizadas por Segmentador, Reconocedor y Clasificador que son controladas desde esta parte de Nessie, son expuestas a continuación en su orden de ejecución.

- Carga del periódico en memoria.
- La imagen de cada página es entregada al Segmentador, para que la divida en noticias.
- Cada noticia es pasada al Reconocedor, que devuelve el texto reconocido en la imagen.
- Los textos de las noticias llegan al Clasificador, para que clasifique la noticia.

De esta manera se clasifica cada una las noticias de los periódicos que se carguen en el sistema.

5. Segmentación de noticias mediante análisis del *layout*.


EL PAÍS, domingo 11 de mayo de 2008 21

ESPAÑA

Dirigentes de IU abren una 'tercera vía' entre la cúpula y los críticos

Un grupo de militantes presenta una candidatura alternativa en Euskadi

Varios dirigentes han suscrito el manifiesto 'Creemos en el futuro de IU'

El PSOE acomete una profunda reforma de su funcionamiento

El congreso socialista flexibilizará las formas de militancia

Grandes ciudades

Si queremos que el usuario pueda acceder directamente a sus noticias de interés y sólo a estas, debemos primero detectar todas las noticias y separarlas unas de otras. Esta es la función principal de este módulo: segmentar cada una de las páginas en las noticias que contiene. Y es el primer módulo de Nessie que entra en acción.

Para desarrollar esta función, destacamos dos partes principales: el Segmentador y el *Article tracker* (rastreador de noticias).

Pero antes de entrar a detallar estas partes, hay que mencionar que se realizan unas tareas previas con el objetivo de mejorar el comportamiento del módulo. Se limpia la página escaneada del ruido que contenga (ruido por ejemplo, son las manchas que pudieran aparecer durante su inicial impresión, su posterior conservación o durante el escaneado).

Esta fase de pre-procesamiento también incluye otras acciones que permitirán que la página sea segmentada con mayor eficiencia y en menor tiempo.

El Segmentador, realiza la tarea de dividir cada página en bloques más elementales y clasificarlas como un bloque de texto, una imagen, un pié de foto o un encabezado, entre otros tipos. Estos bloques, deben ser detectados individualmente, para poder procesar su contenido el módulo de Extracción de texto (explicado en la siguiente sección).


EL PAÍS, domingo 11 de mayo de 2008 21

ESPAÑA

Dirigentes de IU abren una 'tercera vía' entre la cúpula y los críticos

Un grupo de militantes presenta una candidatura alternativa en Euskadi

Varios dirigentes han suscrito el manifiesto 'Creemos en el futuro de IU'

El PSOE acomete una profunda reforma de su funcionamiento

El congreso socialista flexibilizará las formas de militancia

Grandes ciudades

El *Article tracker*, deberá identificar cuáles de los bloques obtenidos anteriormente son titulares o forman parte de uno. Una vez identificados los titulares, podremos identificar y situar las noticias que existen en la página. Cada bloque segmentado se va relacionando con el titular que con mayor probabilidad le corresponda. Esta asignación se hace en base posicionamiento relativo entre el titular y dicho bloque de texto, y teniendo en cuenta elementos separadores que pudieran existir entre ellos.

Tras este proceso ya tenemos los bloques elementales de cada página agrupados por noticia. Así que sólo falta emplear las coordenadas de los bloques que están más a los extremos, para delimitar el área que contendrá la noticia. Esta región será la que finalmente se le mostrará en Nessie al usuario que haya indicado interés por su temática.

Pero para poder llegar a ese punto, son imprescindibles dos tareas: reconocer el texto y clasificar la noticia. Como comentamos, cada bloque sería pasado al Reconocedor de texto de Nessie (o Nessie OCR). Pero además, con cada noticia reconocida por el Segmentador, hay que realizar su clasificación en base al texto contenido en todos los bloques que la forman. Esta tarea la realizará el Clasificador de noticias de Nessie. Ambos módulos se explican en las dos siguientes secciones del documento.

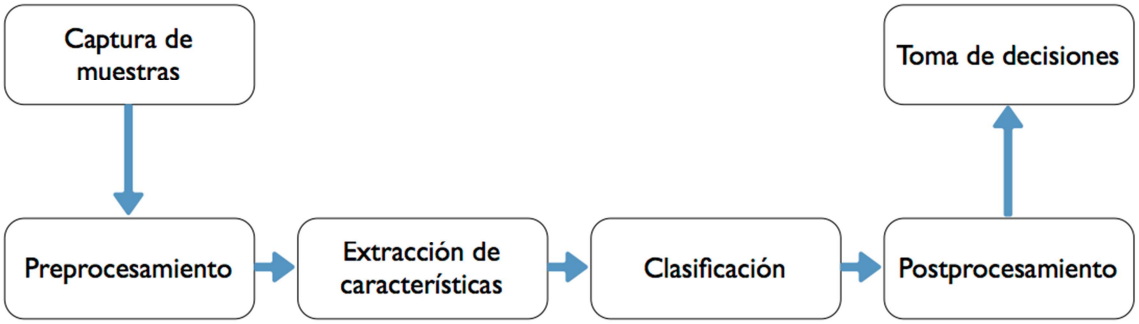
6. Extracción de texto con técnicas OCR.

Cualquier intento de automatizar la "lectura" de un texto a través de una aplicación informática conlleva aplicar de un modo u otro, técnicas de reconocimiento óptico de caracteres (OCR). A fin de cuentas es el mecanismo que permite convertir las imágenes en palabras, rescatando el contenido que verdaderamente interesa al cliente.

El reconocimiento óptico de caracteres es el resultado de aplicar de manera específica técnicas de clasificación de patrones, una disciplina que representa un eje fundamental dentro de la inteligencia artificial y la visión por computador. Identificar patrones diferenciando unos de otros es una tarea que el ser humano tiene perfectamente automatizada, a través de su compleja red sensorial, sus mecanismos de abstracción y su extraordinaria capacidad de razonamiento. Gracias a ellos hemos conseguido sobrevivir a lo largo de toda nuestra existencia. Dotar a una máquina de ese comportamiento constituye una tarea cuanto menos ambiciosa y compleja, pero que en un dominio más reducido sí pueden ser imitados.

Dentro del proyecto Nessie, el módulo de reconocimiento de caracteres se inserta justo en el medio de una secuencia de tres pasos. Recibe como entrada un conjunto de imágenes que representan bloques de texto pertenecientes a una noticia. Tras procesarlos produce un conjunto de textos equivalentes que alimentan una posterior fase de clasificación temática.

En cierto modo podemos decir que de la misma manera que la segmentación de noticias descompone una página en bloques de noticias, un OCR descompone un texto en sus diferentes caracteres. En el caso de Nessie esto se ha logrado en un proceso de seis etapas:



En la captura de muestras la dificultad reside en superar las limitaciones de calidad del escáner que captura las páginas del periódico.

En la etapa de pre-procesamiento se aplican una serie de transformaciones a las imágenes para poder separar los caracteres del fondo, del ruido o de otros caracteres anexos.

La extracción de características, se encarga de seleccionar las propiedades de la forma de las letras, que más permiten distinguir a unas de las otras.

El trabajo de un clasificador de patrones consiste en evaluar el conjunto de propiedades extraído y decidir de qué carácter se trata.

El post-procesado de los bloques de noticias genera meta-información sobre el texto: número de caracteres, tamaños de fuente encontrados, grosor de la tipografía empleado,...

La toma de decisiones se delega en el Controlador que mencionamos previamente. A continuación veremos el módulo de Clasificación, el cual realiza sus acciones tras haberse reconocido todo el texto de las noticias, y con esa información poder clasificarlas.

7. Clasificación de noticias por coincidencia temática.

Una vez explicados los subsistemas de Segmentación y Reconocimiento, queda explicar el tercer subsistema: El Clasificador de noticias en temas. Se encarga de realizar una clasificación temática de cada noticia a partir de un conjunto de palabras clave.

Cada tema tiene una serie de palabras clave asociadas. Por lo que la clasificación se hace de tal manera que si en el texto de las noticias aparece alguna palabra clave de un determinado tema, la noticia se asocia a este.

Esto puede ocasionar que se clasifiquen noticias en temas incorrectos, debido a que esa palabra clave en el contexto de la noticia no debiera ser clasificada en ese tema. Por ejemplo, si usamos la palabra “bolsa” como palabra clave para el tema “Economía”, podríamos obtener como resultado la clasificación dentro de dicho tema noticias que no tienen nada que ver.

Por lo tanto, la elección de las palabras clave, que son definidas por el operario de la aplicación Nessie, debe ser lo más certera y completa posible para aumentar el acierto de la clasificación.

8. Interfaz web.

A través de la interfaz web los distintos usuarios con distintos roles pueden realizar tareas de carga de periódicos a clasificar, administración del sistema, validación de resultados y visualización de noticias clasificadas. Estas dos últimas, son las que procederemos a explicar, ya que son las más llamativas y aportan mayor valor a la interfaz.

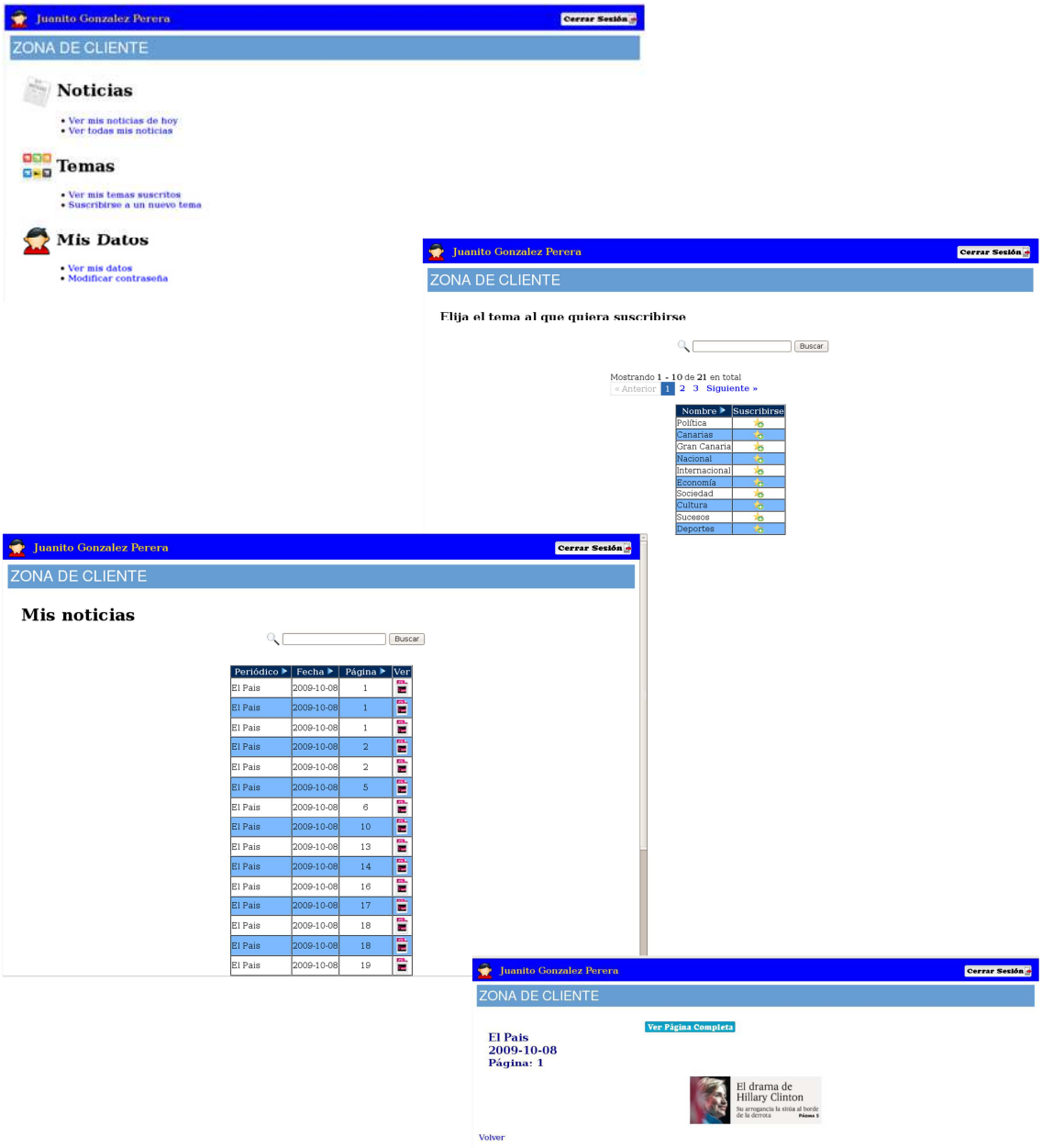
Validación de resultados

Una vez que se ha finalizado con el proceso de clasificación, una persona debe validar los resultados obtenidos. A ese proceso se le ha llamado "Validación de resultados". Tiene la funcionalidad principal de validar los resultados obtenidos en el proceso de segmentación y clasificación.

Se muestra la imagen del periódico, separándose las noticias por color. Cada noticia está formada por una o más áreas. Mediante la manipulación de estas áreas, el operario podrá corregir, si fuera necesario, la división en noticias realizada por el módulo Segmentador. Por otro lado, la interfaz permite en este punto, modificar el tema que fue asociado por el Clasificador, cuando fuera oportuno.

Noticias clasificadas

El cliente final, cuando se autentifica en la aplicación, puede declarar o modificar sus temas de interés, así como ver las noticias que han sido clasificadas en dichos temas.



9. ¿Dónde estamos y hacia dónde vamos?

Este proyecto ha sido una experiencia introductoria para todos en este mundo del seguimiento de medios, así como en muchos de los conocimientos técnicos empleados para su elaboración. Y si además tenemos en cuenta los buenos resultados obtenidos, podemos decir que Nessie tiene mucho potencial. Y que es capaz de solventar muchos de los principales problemas detectados en el rastreo de noticias en prensa escrita, que se realiza regionalmente.

	Tiempos de ejecución (por página)	Acierto en ...
Proceso manual	**4 minutos**	**¿?**
Segmentador de páginas	6,5 segundos	3 de cada 4 noticias
Reconocedor de texto	27 segundos	4 de cada 5 caracteres
Clasificador de noticias	1 segundo	9 de cada 10 páginas
Nessie	**35 segundos**	**3 de cada 4 noticias**

Nessie procesa cada página en poco más de medio minuto, a lo que hay que añadirle el coste aún necesitando del proceso manual de escaneado de los mismos y su carga en la aplicación.

Partiendo de una base de 80 páginas por periódico, tenemos que la aplicación tarda en realizar la carga del periódico algo menos de 7 minutos, a los que hay que sumar el tiempo por página comentado. **Con lo que entre 50 y 55 minutos, se habrá terminado algo que antes, en su elaboración manual, requería 4 horas**. Cabe mencionar, que estos datos fueron obtenidos en ordenadores de gama media y sin explotar las posibilidades de procesado simultáneo de diferentes periódicos.

Por lo tanto, Nessie permite reducir a una cuarta parte el coste temporal (y todo lo que conlleva) el procesado de los periódicos. O dicho de otra manera, aumentar el rendimiento de cada documentalista hasta un 400%.

Pero no todo queda aquí, pues pensamos que Nessie, como todo en la vida, es mejorable. Y cada una de las tareas que son realizadas, son susceptibles de investigación en nuevas técnicas con las que podamos aumentar la tasa de acierto, o mejorar los tiempos de ejecución. Lo que haría a Nessie más robusto y más veloz.

Por todo esto, nos alegra eliminar la etiqueta de leyenda, y poder avistar a Nessie como una realidad.