ESTUDIO SOBRE LA INFORMACIÓN DE TEXTO CONTENIDA EN IMÁGENES WEB

Memòria del Projecte Fi de Carrera
d'Enginyeria en Informàtica
realitzat per
Sergi Robles Mestre
i dirigit per
Dimosthenis Karatzas
Bellaterra, 26 de juliol de 2009

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN

1.1. El Problema

La indexación y la búsqueda de páginas WWW se basa en el análisis del texto. La tecnología actual, aún no puede procesar de una manera eficiente y suficientemente rápida el texto contenido en las imágenes de las páginas WWW. Este hecho plantea un problema importante ya que el texto en forma de imagen suele ser significativo semánticamente (por ejemplo encabezados, títulos, anuncios, etc.).

Otro problema es el de la accesibilidad, especialmente para personas ciegas o para otros formatos compactos de Internet que no muestran las imágenes (navegadores de sólo texto/voz.. etc)

Estos dos grandes problemas del uso de imágenes con contenido de texto: indexación / búsqueda y accesibilidad, se agravan si el texto contenido en esas imágenes no aparece en ningún otro sitio de la página.

El uso de imágenes es de lejos la manera más común de añadir contenido audiovisual en los documentos de texto plano. Pero las imágenes no son usadas sólo para ilustraciones sino que en muchas ocasiones contienen texto. El texto en las imágenes puede tener varios fines como el de añadir un impacto a un mensaje de texto, mostrar palabras de un idioma que no se puede representar de forma estándar con HTML o mostrar operaciones matemáticas. Otras veces el único fin de la imagen es mostrar un tipo de fuente no estándar que el diseñador quiere utilizar.

El HTML proporciona una descripción alternativa textual para las imágenes utilizando el atributo ALT. En la actualidad también puede utilizarse el atributo TITLE ya que con la aparición del DHTML (Dynamic HTML) y del CSS no sólo los elementos que son del tipo IMG pueden tener imágenes (la propiedad ALT es especifica del objeto IMG). Cualquier elemento: capa, botón, encabezado puede mostrar imágenes de fondo mediante CSS.

Si el texto descriptivo que se encuentra en los atributos ALT y/o TITLE correspondiera siempre con el texto contenido en las imágenes el problema sería menos importante pero la realidad muestra que esto no es así. Varios estudios [1][2] nos muestran que en una gran parte (sobre el 60%) de éstas etiquetas no se incluyen o se incluyen con una descripción falsa que no coincide con el texto que contiene la imagen.

Estos problemas de indexación e inaccesibilidad no son sólo producidos por las imágenes; la web ha evolucionado en los últimos años y ahora permite una gran cantidad de contenidos audiovisuales extra como películas de Flash, vídeos multimedia, documentos PDF. Todos éstos formatos tienen los mismos problemas que las imágenes; la información textual que contienen es ignorada por los buscadores y invisible para los navegadores de texto/voz.

Presentación

A continuación se expondrán los objetivos de este proyecto y se presentará tanto una planificación para dicho proyecto como un plan de viabilidad.

Más adelante se mostrará el Estado del Arte sobre este problema y se explicarán las características del software que se ha desarrollado.

Después se analizarán y comentarán los resultados obtenidos del estudio sobre la situación actual de Internet.

Para concluir se expondrán las conclusiones del trabajo así como los objetivos conseguidos y el trabajo futuro.

1.2. Objetivos

Queremos ver en qué relación el texto que contienen las imágenes en el WWW corresponde con su texto descriptivo o bien aparece en algún otro sitio de la página. Este estudio debe ser lo más representativo posible de la situación actual en la web. A parte queremos que el estudio se pueda hacer de una manera lo más automática posible para ganar en velocidad y simplicidad y permitir que se pueda volver a realizar siempre que se quiera para ir controlando la evolución. Para ello deberemos desarrollar una aplicación software.

Una vez se tengan los resultados de éste estudio deberán ser analizados, comparados con estudios anteriores [1][2] y sacar conclusiones.

Los objetivos específicos para el proyecto son:

O1. Escoger un conjunto de páginas que sean representativas del estado actual en la Web.

O2. Diseñar una base de datos para almacenar toda la información significante de las páginas web analizadas.

O3. Implementar un software que ayude, guiando al usuario, al análisis y a la clasificación de las páginas y a las imágenes de éstas. Las funcionalidades más específicas que deberá cumplir el software se explicarán más adelante en el apartado de desarrollo del software, pero aquí adelantamos los 3 objetivos básicos que deben realizarse:

- **O3.1. Extracción de Información de manera automática** de una página web, como texto e imágenes y propiedades de estas (tipo, dimensiones, bpp).
- **O3.2. Introducción manual de Información**. Deben poder clasificarse las imágenes según si tienen texto o no y en el caso de que tengan texto habrá que poner la transcripción de ése texto.
- **O3.3. Sacar resultados de las páginas analizadas**. Permitir crear varias selecciones de páginas y con ellas analizar las características que nos interesan.

O4. Analizar la situación actual de la web. Mediante el software desarrollado deberá analizarse el conjunto de páginas representativas del estado actual en la Web y sacar los siguientes resultados:

- **O4.1.** Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.
- **O4.2.** Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.
- **O4.3.** Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.
- **O4.4.** Porcentaje del área de la página cubierta por imágenes que contienen texto.
- **O4.5.** Comparar los resultados con un estudio [1] del año 2001.

2. Planificación y viabilidad

La realización del proyecto fue planificada desde su principio (diciembre 2008). La planificación contempla tanto los plazos de realización de varias tareas así como las fechas de entrega de informes y solicitudes.

Una vez por semana se ha realizado una reunión para verificar que la planificación seguía correcta y no se debían realizar ajustes. Según la tarea que se estuviera realizando, en la reunión se hacía una revisión más específica de los objetivos de la tarea en cuestión para poder cumplir con su "deadline".

Las tareas que se planificaron son:

-**Obtención de páginas web**. Realizar un estudio sobre Internet, escoger un conjunto de páginas representativo y descargar las páginas

- **Informe Previo**. Redacción del informe previo.

- **Crear la Base de Datos** que utilizará el software a desarrollar.

- **Desarrollar la aplicación software**. Debido a la complejidad de ésta tarea, se dividió en subcategorías:

    -**Extracción automática de Información**. El programa debe extraer toda la información que pueda automáticamente de una página web; texto de la página, imágenes, propiedades de las imágenes (dimensiones, formato, bpp, posición en la página..)

    - **Introducción manual de Información.** Creación de una GUI (Graphics User Irnterface) donde se vean las imágenes de la web y se puedan clasificar según si tienen texto o no y en el caso de que tengan, poder poner la transcripción de ése texto y clasificar la imagen según la función que realice.

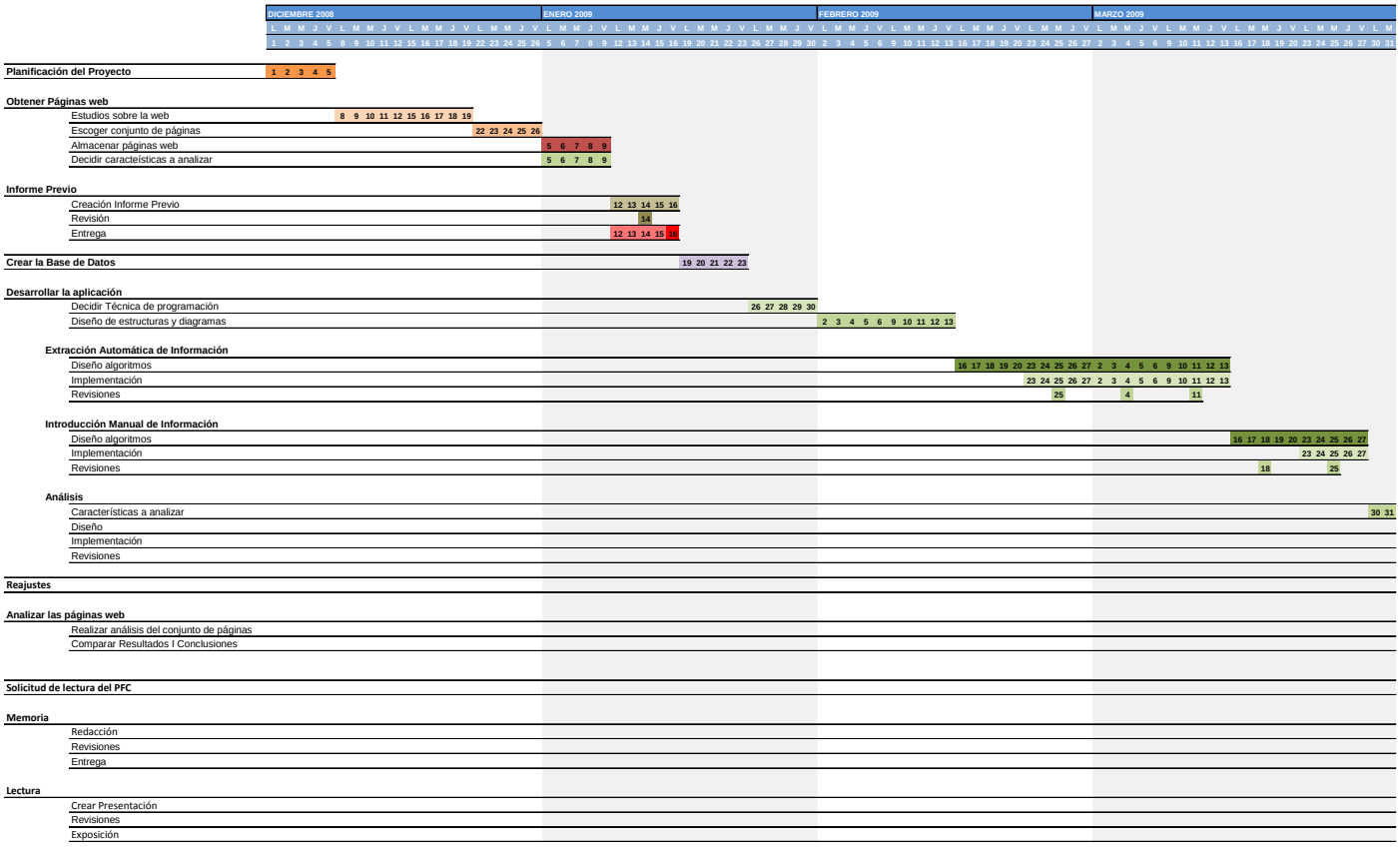
    - **Análisis.** Extracción de los resultados de las páginas analizadas de una manera gráfica.

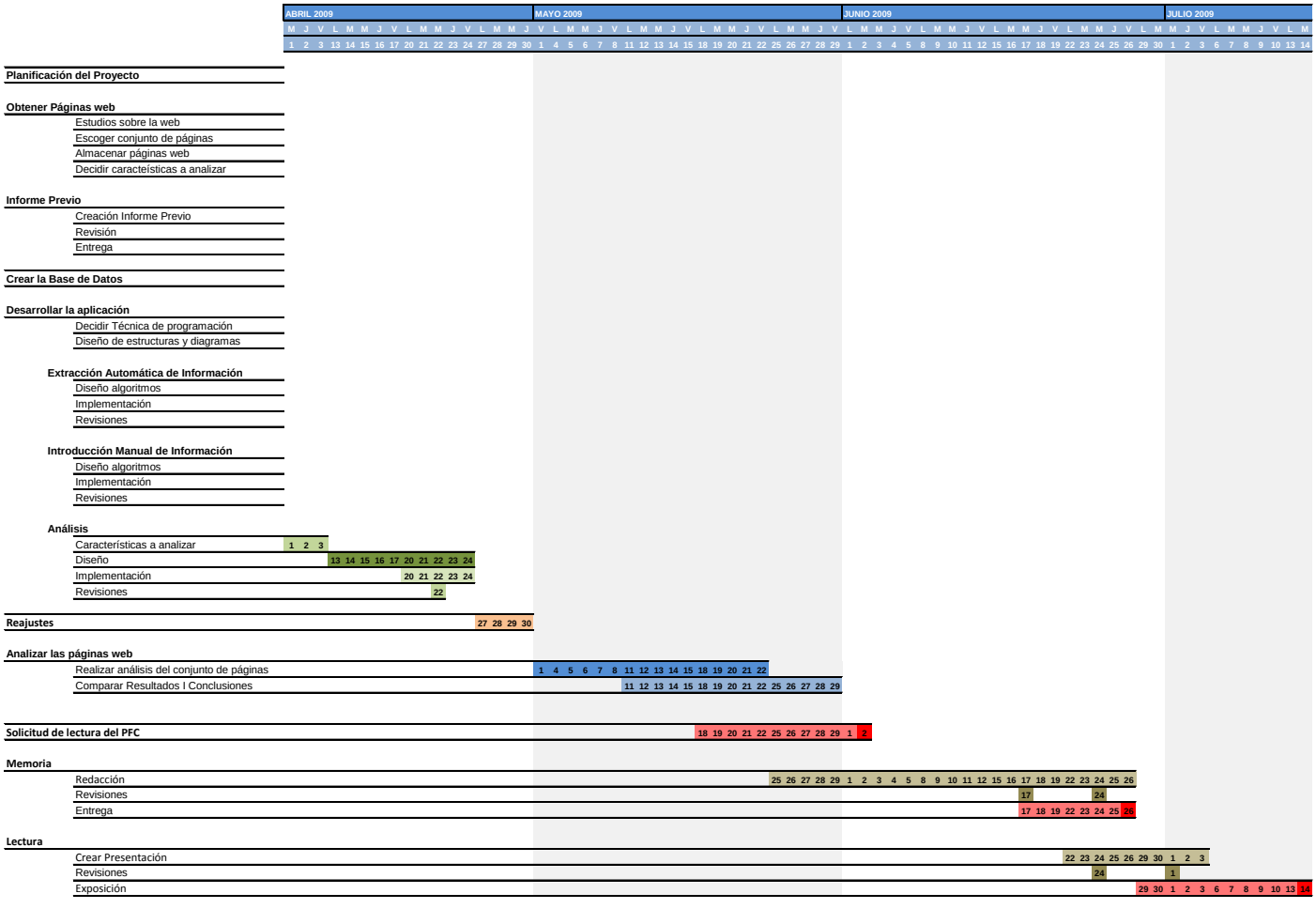
- **Analizar las páginas web** del conjunto de páginas representativo de la situación actual de Internet con la aplicación software ya desarrollada.

- **Memoria**. Redacción y revisiones de la Memoria.

-**Lectura**. Crear la presentación para la Lectura. Preparar la Lectura.

Según todas estas tareas se creó la siguiente planificación que ya fue incluida en el informe previo:


DICIEMBRE 2008
ENERO 2009
FEBRERO 2009
MARZO 2009
Planificación del Proyecto
Obtener Páginas web
Estudios sobre la web
Escoger conjunto de páginas
Almacenar páginas web
Decidir caracteísticas a analizar
Informe Previo
Creación Informe Previo
Revisión
Entrega
Crear la Base de Datos
Desarrollar la aplicación
Decidir Técnica de programación
Diseño de estructuras y diagramas
Extracción Automática de Información
Diseño algoritmos
Implementación
Revisiones
Introducción Manual de Información
Diseño algoritmos
Implementación
Revisiones
Análisis
Características a analizar
Diseño
Implementación
Revisiones
Reajustes
Analizar las páginas web
Realizar análisis del conjunto de páginas
Comparar Resultados I Conclusiones
Solicitud de lectura del PFC
Memoria
Redacción
Revisiones
Entrega
Lectura
Crear Presentación
Revisiones
Exposición


ABRIL 2009
MAYO 2009
JUNIO 2009
JULIO 2009
Planificación del Proyecto
Obtener Páginas web
Estudios sobre la web
Escoger conjunto de páginas
Almacenar páginas web
Decidir caracteísticas a analizar
Informe Previo
Creación Informe Previo
Revisión
Entrega
Crear la Base de Datos
Desarrollar la aplicación
Decidir Técnica de programación
Diseño de estructuras y diagramas
Extracción Automática de Información
Diseño algoritmos
Implementación
Revisiones
Introducción Manual de Información
Diseño algoritmos
Implementación
Revisiones
Análisis
Características a analizar
Diseño
Implementación
Revisiones
Reajustes
Analizar las páginas web
Realizar análisis del conjunto de páginas
Comparar Resultados I Conclusiones
Solicitud de lectura del PFC
Memoria
Redacción
Revisiones
Entrega
Lectura
Crear Presentación
Revisiones
Exposición

La planificación ha sido realista y se han podido cumplir la mayoría de plazos de las distintas tareas. La única excepción fue la tarea de desarrollo de software que duró dos semanas más de lo previsto inicialmente debido a que nos encontramos con varios problemas de diseño que comentaremos más adelante (apartado 4.2). Cómo en la planificación ya incluíamos una semana para reajustes, en total sólo perdimos una semana. Hemos recortado una semana la duración de la tarea de Analizar las páginas web.

El proyecto ha sido viable: se ha cumplido con la planificación y se han logrado los objetivos de éste con los recursos humanos y tecnológicos que se requerían. Estos recursos eran:

- Herramientas Hardware
    - Un ordenador con conexión a Internet.
- Herramientas Software
    - Máquina Virtual de Java.
    - El sistema operativo es indiferente: Windows / Linux ya que el software se implementará en Java.
    - Navegador Web Firefox.
    - Base de Datos MySQL.
- Recursos Humanos
    - 1 persona durante 6 meses.

3. Estado del Arte

3.1 Revisión de literatura

Los problemas de indexación y accesibilidad del texto contenido en imágenes en las páginas web ya habían sido estudiados en varias ocasiones.

Las reglas de accesibilidad W3C recomiendan que cada imagen asigne un contenido textual equivalente [10].

En una publicación [1] del año 2001 vemos que estas reglas no se cumplen en su gran mayoría; el 56% de las etiquetas ALT en las imágenes con texto eran falsas, incompletas o no existían. Otro dato importante era que la mayoría (76%) de las palabras que aparecían en las imágenes no aparecían en el texto. Cabe destacar que para éste estudio las 200 páginas analizadas no respondían a ningún criterio especial y muchas de ellas eran de las mismas temáticas; los resultados podrían ser algo distintos si se hubiera tenido en cuenta la situación que había en la web en ese momento.

Estos datos fueron confirmados por otra investigación [2] del propio año 2001 y se indicaba que el uso de imágenes con texto estaba aumentando.

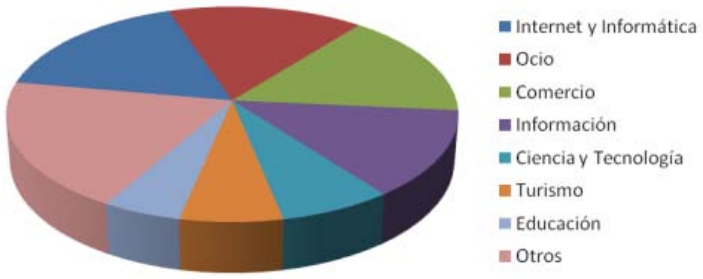
Más adelante en el año 2006, otra publicación [3] decía que la magnitud de éste problema seguía siendo prácticamente la misma: sólo el 39,6% de las imágenes poseían el texto alternativo. En esta publicación [3] se presentaba un sistema para etiquetar automáticamente imágenes que no disponen de texto alternativo. Para ello el sistema se basa en el contexto de la imagen en la página web, en un OCR automatizado y en caso de no funcionar en el etiquetado humano.

Se han estudiado varios métodos para el reconocimiento del texto contenido en las imágenes pero ninguno de estos métodos funciona siempre para todas las imágenes. Para una imagen (o una zona de la imagen) hay siempre un método más adecuado que otro. Una publicación [1] explicaba un método de reconocimiento de texto en imágenes basado en la segmentación de las imágenes y en cambios de iluminación.
En otra publicación [4,5,6] para extraer texto de imágenes complejas en la web primero cuantificaban el color y luego detectaban los componentes conectados.
Otra propuesta [7] se basa en un algoritmo basado en analizar la textura de la imagen.

También se han presentado algoritmos [8,9] para detectar el texto en secuencias de imágenes de video.

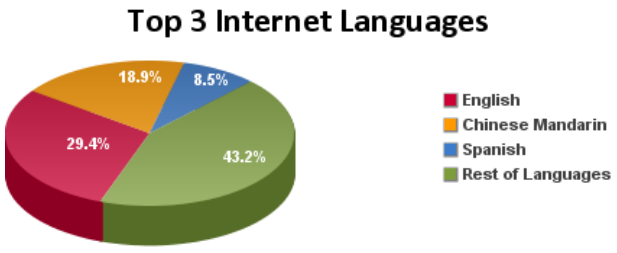
3.2 Análisis de la situación actual de la web.

Para que los resultados de nuestro estudio sean lo más representativos posible de la situación actual de la web habrá que escoger de forma adecuada el conjunto de páginas a analizar.
Las características que hemos escogido de cada página son la categoría a la que pertenece (información, ocio, internet..) y el idioma.
La clasificación de las páginas por categorías puede hacerse según varios criterios: categorías de más interés, categorías más visitadas, o clasificación según las categorías existentes.
Es difícil encontrar estudios que atiendan a la clasificación de páginas según su categoría sobretodo de las categorías existentes.
Hemos combinado un estudio de categorías de más interés [11] con otro de categorías existentes [12] para crear la siguiente clasificación:



Las categorías más importantes según el estudio son Internet e Informática, Ocio, Comercio e Información. Estas categorías tienen unos pesos más importantes comparadas con otras como Turismo y Educación.

La clasificación por idiomas [13] de las páginas web es la siguiente:



Vemos que el inglés es la lengua más utilizada seguida del chino y el español.

Nuestro estudio deberá tener en cuenta estas dos clasificaciones.

4. Desarrollo de Software

Hemos desarrollado una aplicación software para poder automatizar lo más posible el proceso de creación del estudio del problema que nos concierne. En otras publicaciones [1] éste estudio se realizaba de forma completamente manual, analizando cada página, contando palabras e imágenes y al final calculando porcentajes de todas las características.

Nuestra aplicación permitirá simplificar al máximo el proceso, requiriendo únicamente la interactuación humana en aspectos que son imposibles de automatizar (decidir si la imagen tiene texto o no y en caso afirmativo especificar la transcripción del texto).

A continuación explicaremos en detalle todas las funcionalidades que debe tener el software desarrollado. Teniendo en cuenta estas funcionalidades explicaremos con qué tecnologías se ha desarrollado y el porqué.

Más adelante daremos una visión de la estructura del software y explicaremos su uso.

4.1 Funcionalidades

La aplicación deberá realizar las siguientes funciones:

- Extracción automática de Información. Debe extraerse toda la información que se pueda de forma automática de una página web. La información de una página web que nos interesa conocer es:

- **Titulo de la página**

- **URL**

- **Texto que contiene la página**. Nos interesa todo el texto que aparece escrito en la pagina, el visible y el no visible ya que estos, junto con el texto que figura en los meta tags y el propio título de la pagina es el que indexan los buscadores. Para conseguir este texto de forma automática se deberá analizar el código fuente de la página y eliminar todos los tags de HTML y el texto que figure dentro de tags como el <script> </script>

- **Texto de los meta tags**. Nos interesa el texto que aparece en los meta tags description y words.

- **Texto que aparece en las propiedades ALT/TITLE**.

- **Dimensiones de la página**. Queremos saber las dimensiones en píxeles de la página web para luego poder realizar análisis como el de ¿qué porcentaje de cobertura tienen las imágenes en la página?

  Para que tenga sentido éste análisis deberemos fijar el ancho de cualquier página antes de analizarla, un ancho de 1024 píxeles es adecuado.

- **Imágenes en la página**. Cualquier imagen que aparezca en la página deberá ser extraída. De las imágenes también queremos saber algunas propiedades:

  - **Dimensiones** de la imagen (ancho x alto).

  - **Formato** de la imagen (JPG, GIF, PNG…)

  - **BPP**. Bits Por Pixel de la imagen.

- **Referencias a las imágenes**. Cada imagen puede aparece en la página web una o más veces. Para cada referencia de la imagen necesitaremos conocer cierta información como:

- Posición. Coordenadas (superior,izquierda) de la referencia a la imagen.

- Tamaño de la referencia a la imagen. Éste tamaño no siempre es igual al tamaño real de la imagen, puede ser superior o inferior. Si la referencia a la imagen es del tipo IMG la imagen puede aparecer escalada si se han especificado los atributos **width** y/o **height.** Si la referencia es de otro tipo la imagen aparece como un fondo y puede visualizarse sólo una parte de ella, o toda ella repetida en la dirección del eje x y/o del eje y.

Ejemplo de tamaño de referencia inferior al de la imagen.

Imagen Original	
Ancho	168px
Alto	119px

Referencias			
Ancho	18px	16 px	66 px
Alto	26 px	26 px	26 px
Offset - X	-26 px	-44 px	-168 px
Offset - Y	0 px	0 px	0 px

El fondo (imagen) se mueve según el offset para mostrar sólo la parte que se necesita.

Ejemplo de tamaño de referencia superior al de la imagen.

Imagen Original	
Ancho	1px
Alto	835px

Referencia	
Ancho	1007px
Alto	1230px
Background-repeat	Repeat-x

La imagen se repite a lo ancho del eje x , ya que así lo indica su propiedad background-repeat (repeat-x)

Para poder mostrar correctamente la referencia de una imagen, a parte del tamaño, según lo que acabamos de mencionar, necesitaremos saber otras propiedades adicionales como la **background-position** (offset de las coordenadas x e y) y **background-repeat** (no-repeat, repeat-x, repeat-y, repeat).

- Visibilidad. Queremos saber si la referencia es visible o no en la página. Cada objeto en HTML tiene una propiedad que indica si el objeto es visible o no, pero para que el objeto sea visible en la página, además de que su propiedad de visibilidad sea cierta también deben serlo las de los objetos que lo contienen (objetos padre).

- Z-Index. Nivel de profundidad de la referencia. Según este valor las referencias se sitúan unas encima de otras.

- Tipo de etiqueta. Nos indica de qué tipo de objeto (IMG,DIV, BUTTON, INPUT…) es la referencia a la imagen.

- Introducción manual de Información. Creación de una GUI (Graphics User Irnterface) donde se vean las imágenes de la web y se puedan clasificar según si tienen texto o no. Si tienen texto se deberá permitir:

- Escribir la **transcripción** del texto de la imagen.

- Seleccionar la **función** que realiza la **imagen** en la página. Las funciones posibles que hay que mostrar son: Logotipo, Título, Cabecera, Menú-Botó, Advertencia, Subtitulo, Pie, Científica, Contenido Extranjero, Fotografía.

A parte queremos otra información que aunque para éste estudio no la utilizaremos podrá ser útil en un futuro para el reconocimiento OCR del texto en las imágenes. La información que deberemos poder seleccionar es:

- **Tipo del texto**. Uniforme, Multicolor, Fotográfico, Gradiente, Textura
- **Tipo del fondo**. Uniforme, Multicolor, Fotográfico, Gradiente, Textura, Transparente

- Análisis. Aquí también se deberá crear una GUI en la que se permitan realizar distintas selecciones de páginas y analizar para cada ella las características que nos interesan para el estudio. Las características a analizar son:

- Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.
- Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.
- Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.
- Porcentaje del área de la página cubierta por imágenes que contienen texto.

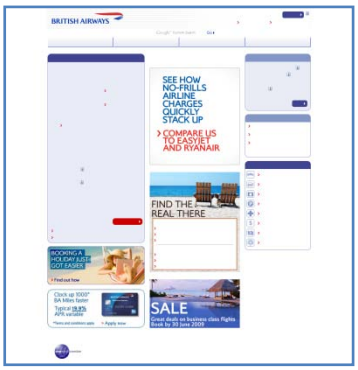
Estás características deberán mostrarse en pantalla de una manera gráfica.

- Vistas de una página web

Deberán mostrarse varias vistas de cada página analizada:

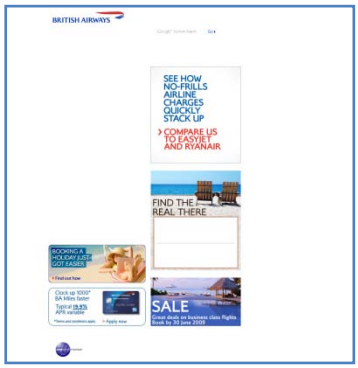
- **Vista de la página sin modificar.**

Es sencillamente una captura de la página web, aunque habrá que ver cómo realizarla ya que la mayoría de páginas tienen un tamaño vertical superior a la resolución del propio monitor, y nosotros queremos una captura completa, no sólo una parte.



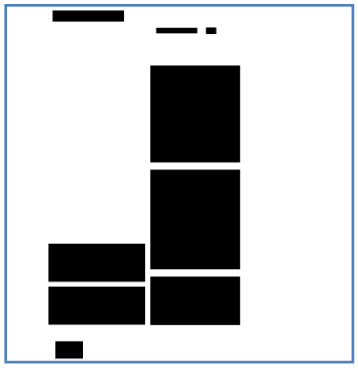
- **Vista de la página sólo con las imágenes.**

Está vista deberá ser construida a partir de todas las imágenes que se han descargado de la página, situándolas en las posiciones adecuadas según la información que también se habrá extraído de la página.



- **Vista de la página sólo con imágenes que contienen texto.**

La misma vista que la anterior pero mostrando sólo las imágenes que contengan texto.



- **Vista de la cobertura de las imágenes con texto.**

Se tendrán que mostrar las imágenes que tengan texto en negro para apreciar claramente la cobertura de éstas en la página. A partir de esta vista podremos calcular el porcentaje de la página cubierto por estas imágenes.

- Configuración

En algún sitio de la aplicación deberán poder configurarse algunas opciones para el correcto funcionamiento de la misma como especificar la ruta del navegador y el servidor de la B.D.

También se deberán poder modificar otras características:

- Categorías de las páginas web y peso de cada categoría.

- Funciones que realizan las imágenes que tienen texto y su peso.

- Tipos de texto de las imágenes que tienen texto.

- Tipos de fondo de las imágenes que tienen texto.

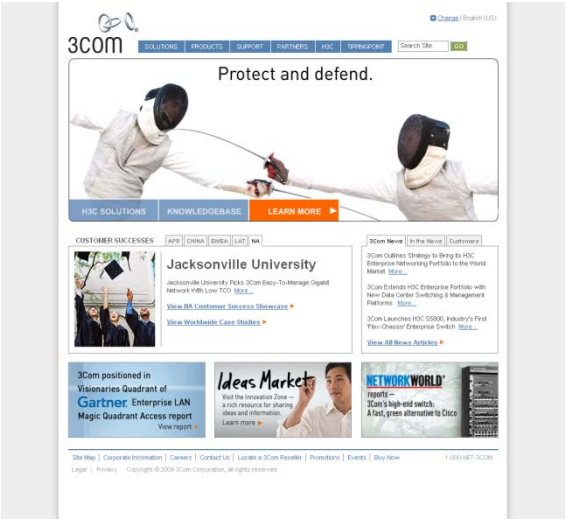
- Listado de palabras que se excluirán del estudio. Cualquier palabra del texto de la página o de las imágenes que aparezca en este listado no será incluida en el estudio.

4.2 Elección de tecnologías

Para el desarrollo del software ya inicialmente se pensó en trabajar con **JAVA** y utilizar una base de datos de **MySQL** debido a la portabilidad de ambos.

Para extraer la información necesaria de las páginas web, primeramente se pensó en descargarlas. Para dicho propósito se había pensado utilizar algún programa libre de descarga de páginas web. Ésta era una opción que años atrás, cuando las páginas web eran más simples (prácticamente sólo existía el lenguaje HTML), hubiera funcionado perfectamente; una página descargada al abrirla en el navegador se visualizaba igual que la original. En la actualidad, debido a la evolución de la web, permitiendo nuevos contenidos (video, C.S.S. y sobre todo JavaScript), el resultado de abrir una página después de descargada en muchas ocasiones no es el esperado.

Captura de la página web de 3COM con conexión a Internet.



Captura de la página web de 3COM después de descargarla.

Se hicieron varias pruebas, incluso se añadió a la aplicación la funcionalidad de descargar la página web. Para esto se descargaba la página deseada y se analizaba el código fuente para encontrar todas las dependencias que tenía (ficheros CSS, JavaScript, imágenes, archivos Flash). Éste proceso era recursivo ya que ficheros JavaScript pueden llamar a otros ficheros Javascript, en los ficheros CSS hay referencias a imágenes, etc.

El resultado final era muy parecido a haber descargado la página con otro programa o hasta con el propio navegador; la mayoría de páginas se veían igual, otras mejor y alguna otra peor.

Para que esta última opción funcionase perfectamente hubiera requerido muchos más recursos tecnológicos y humanos ya que prácticamente se estaría realizando un navegador web que analizase en profundidad todos los ficheros HTML, CSS e JavaScript.

Otra manera de conseguir extraer toda la información de una página web es utilizando el lenguaje **JavaScript**. Éste lenguaje permite (a través del DOM Document Object Model) tener acceso a todos los elementos de una página, consultar y cambiar propiedades. Nos encontrábamos de entrada ante dos grandes problemas:

- Extracción de la información al exterior

Por seguridad el lenguaje JavaScript fue diseñado para que no tuviera acceso al ordenador cliente; las operaciones de entrada/salida no son permitidas. La única operación permitida es la de lectura/escritura en cookies. Hicimos una implementación en la que el código JavaScript escribiá la información en una cookie y luego la aplicación la leía analizando el fichero en el que se guardan las cookies. Aún así quedaba un gran problema:

- Insertar código JavaScript dentro de una página.

Podemos crear el código JavaScript que extraiga la información de la página y hasta la grabe en una cookie pero necesitamos que este código esté dentro de la propia página para que funcione. Esto de entrada es imposible a no ser que se descargue la página y modifiquemos el código fuente de esta, pero en este caso volveríamos al problema anteriormente comentado de descargar las páginas.

La única solución está en tener acceso al propio navegador y añadirle la funcionalidad de extracción de la información que queremos. A parte, si tenemos acceso al propio navegador, también disponemos de acceso a las operaciones de entrada/salida con lo que se podrían grabar ficheros en disco sin tener que utilizar las cookies.

Para poder realizar esto hemos desarrollado una **extensión para el navegador FireFox**. Ésta extensión está programada en **XUL** (XML User Interface Language) e **JavaScript** y funciona tanto para FireFox como para otros navegadores como ThunderBird, SeaMonkey basados en la tecnología Gecko (motor de renderizado).

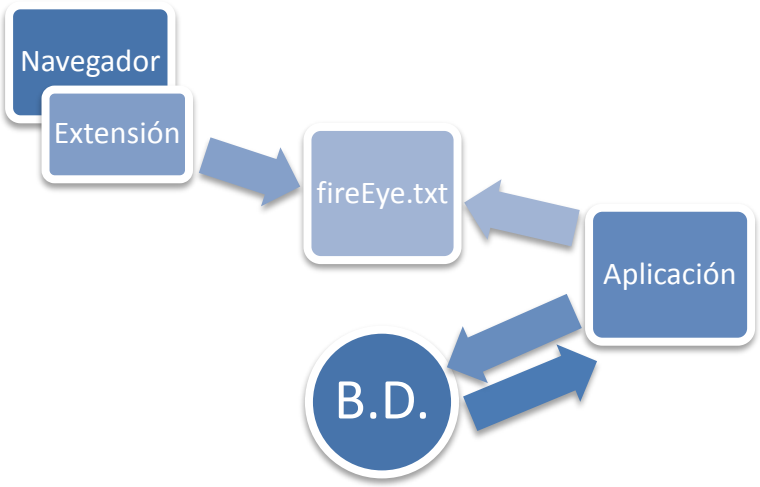
4.3 Estructura

El software desarrollado tiene tres componentes diferenciados:

- Extensión para el navegador (Componente FireEye).
- Aplicación para la gestión de las páginas analizadas y extracción de resultados.
- Base de Datos MySQL.

Comunicación entre componentes.

La extensión para el navegador y la aplicación de gestión se comunican a través de un fichero de texto. La extensión al analizar una página crea un fichero llamado "fireEye.txt" con la información extraída de la página web. En la misma ubicación de este fichero se guardan todas las imágenes de la página y una imagen con la captura de la web. La aplicación leerá este fichero y almacenará en la base de datos toda la información de la página, las imágenes que contiene y su captura.



Componente FireEye.

Éste componente requiere de un navegador que utilice la tecnología Gecko (FireFox, ThunderBird..). El componente se integra al navegador ya que es una extensión.

Este componente es el encargado de extraer toda la información que queremos de una página web. El componente analiza la página mediante JavaScript, explorando el DOM y extrae la información en la ubicación seleccionada. La información que el componente extrae de la página es:

- **Fichero** (fireEye.txt) con información de la página web:
  - **Título**
  - **Dimensiones de la página**

- **Texto que contiene la página**
- **Texto en las etiquetas ALT y TITLE**
- **Texto en las etiquetas META** (description, words)
- **Lista de imágenes que aparecen en la página**.

  Mediante JavaScript se analizan todos los objetos de la página en busca tanto de las propiedades src como background-image que son las que indican que el objeto muestra una imagen. Se crea un conjunto con todos los nombres de las imágenes.
- **Lista de todas las referencias a las imágenes** (cada imagen puede aparecer más de una vez en la página) y las **propiedades** de cada una de ellas: **tamaño**, **posición**, **visibilidad**, **z-index**, **tipo de etiqueta**, propiedad **background-position** y **background repeat**.

  Al igual que para realizar la lista de imágenes se recorre toda la lista de objetos de la página buscando los que tengan imágenes (cómo mínimo habrá tantos elementos como en la lista de imágenes). Para cada objeto se anotan todas las propiedades que queremos:

  Las propiedades tamaño, tipo de etiqueta, background-position y background-repeat se obtienen directamente del objeto.

  La propiedad posición se calcula sumando el offset de los objetos que contienen al objeto actual hasta llegar al objeto principal (root). También deben sumarse el valor de las propiedades margin y border del objeto.

- Imágenes

A partir de la lista de imágenes de la página, estas se pueden grabar a disco gracias a que el componente es una extensión del navegador y como tal puede acceder a funciones propias de este. Sino, como comentamos con anterioridad, en el lenguaje JavaScript esto sería imposible.

- Captura de la página web

Para realizar una captura completa de la página web se añade a la página un objeto HTML de tipo CANVAS con un tamaño igual al área completa de la página. En él se puede "dibujar" una imagen con toda la página web (propiedad drawWindow()). Este objeto lo podemos guardar como imagen en un fichero.

Aplicación de gestión

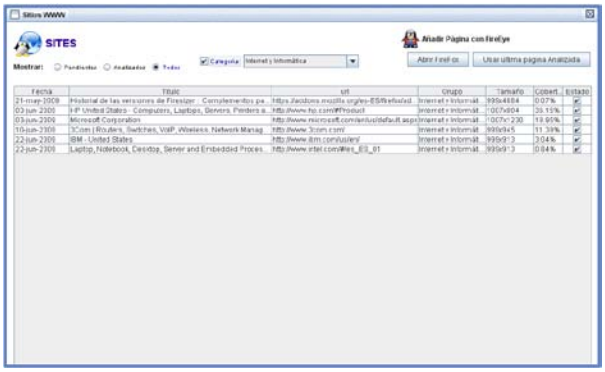
La aplicación se ha implementado en JAVA y se ha puesto una atención especial a que sea fácil de usar y amigable (user friendly). La aplicación permite abrir múltiples ventanas. Los tipos de ventanas que pueden abrirse y las funcionalidades de cada una de ellas son las siguientes:

- **Gestión de páginas**

Muestra un listado con las propiedades básicas de cada página (título, url, dimensiones, página analizada sí o no… etc.). Se pueden aplicar filtros al listado:

- Sólo páginas analizadas.
- Sólo páginas pendientes de analizar.
- Sólo las de la categoría seleccionada.

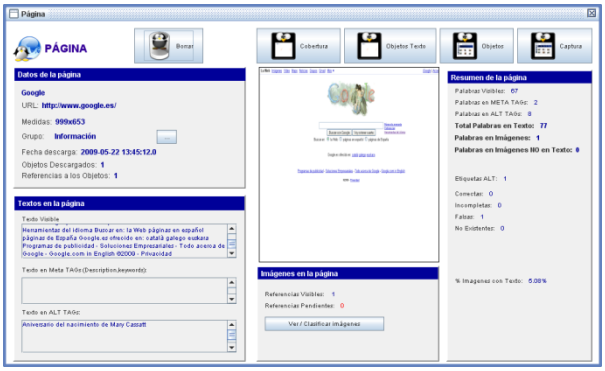
Permite abrir la ficha de la página que se seleccione del listado. También permite dar de alta una nueva página con el componente FireEye.



- **Ficha de una página**

Muestra la información extraída de la página, una captura de ella y un resumen de las características (palabras en imágenes/texto, Etiquetas ALT correctas/incorrectas…) del estudio para la página.

Tiene unos botones para acceder a la ventana de vistas de la página web y un botón para borrar la página.

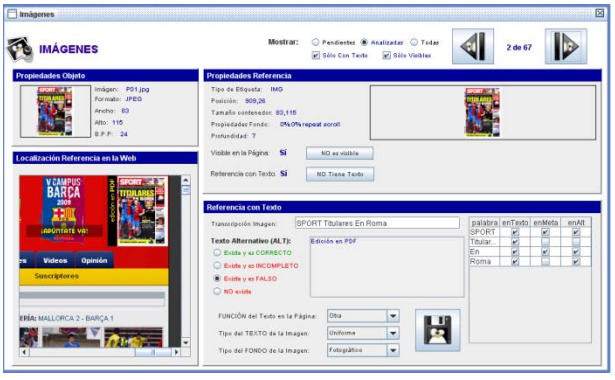


- **Imágenes**

En esta ventana se muestra una a una cada referencia a imagen que aparece en la página web para que sea clasificada según si tiene texto o no.

Se puede ver y ampliar la referencia y también la captura de la página web con la localización de la referencia.

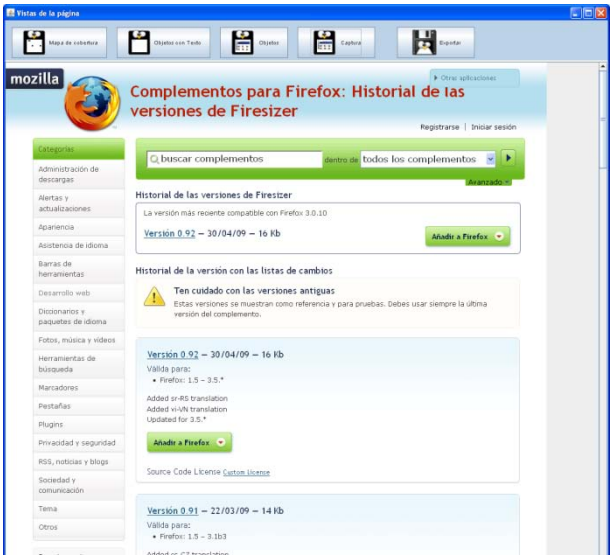
En el caso que la imagen tenga texto, se deberá introducir más información como la transcripción del texto, la función que realiza la imagen en la página y el tipo de texto y de fondo de la imagen.



- **Vistas de la página**

En esta ventana se muestran cuatro posibles vistas de una página analizada; la propia captura, una versión de la página con sólo las imágenes, otra con sólo las imágenes que contienen texto y una última que es el mapa de cobertura de las imágenes con texto de la página.

Cualquier vista de las cuatro puede ser exportada a un fichero.

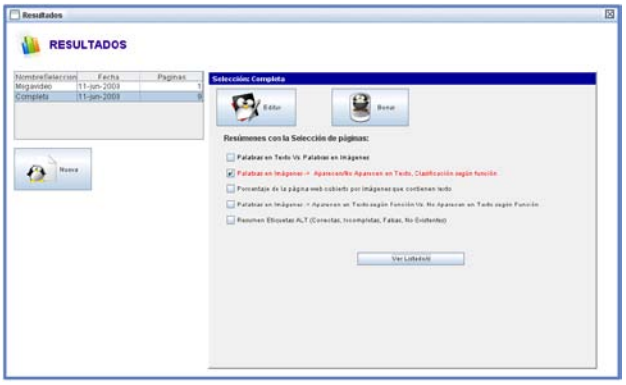


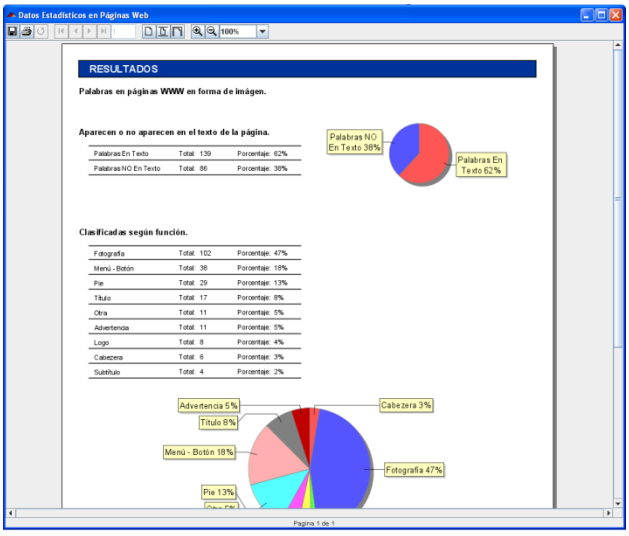
- **Resultados**

Se permiten realizar diversas selecciones de páginas y con estas selecciones mostrar los resultados de las características que se deseen analizar; estas eran:

- Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.
- Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.
- Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.
- Porcentaje del área de la página cubierta por imágenes que contienen texto.

Para mostrar los resultados de forma gráfica se ha incluido en la aplicación la librería JasperReports.

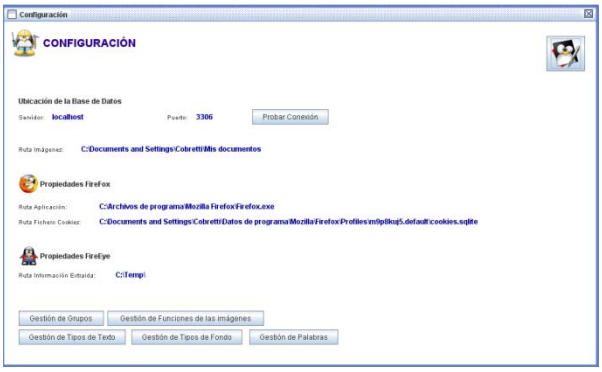




- **Configuración**

En esta ventana se configuran todos los parámetros importantes de la aplicación como las rutas de la B.D., del programa FireFox y del componente FireEye.

También se configuran otros aspectos como las categorías de las páginas, funciones de imágenes, tipos de texto y fondo en las imágenes y palabras que no se contemplarán para el estudio.



Base de Datos MySQL

Se creado una base de datos en MySQL para almacenar y gestionar toda la información que se extrae de las páginas web. Cabe destacar que se incluyen todas las imágenes de las páginas y las capturas de estas.

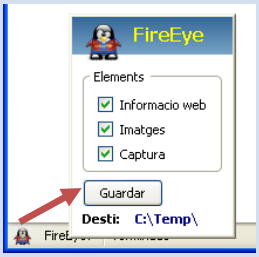
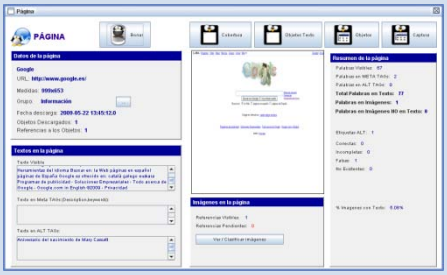
No entraremos en detalle sobre la implementación o diseño, tan sólo comentar que la base de datos tiene definidas todas las restricciones necesarias entre tablas y tiene procedimientos almacenados, triggers y vistas necesarios.

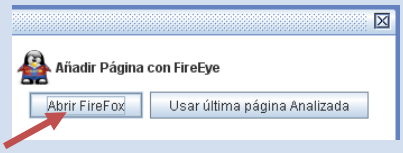
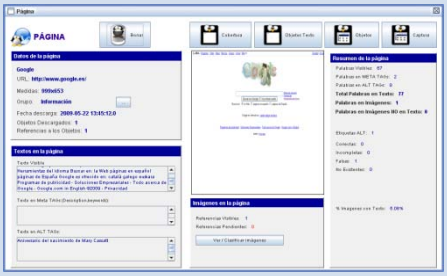
4.4 Uso del software

A continuación expondremos la manera en que debe usarse el software desarrollado, explicando primero las opciones para analizar una página web y a continuación como proceder para clasificar las imágenes que aparecen en las páginas analizadas. Para concluir explicaremos como visualizar los resultados de las páginas analizadas.

-Analizar una página web.

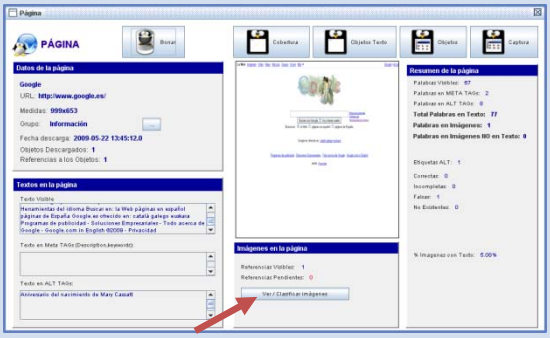
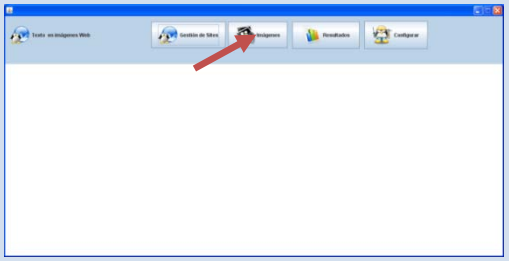
El análisis de una página web se realiza desde el componente FireEye que hemos desarrollado. Eso quiere decir para analizar una página debemos tener abierto el navegador con esa página. Tenemos dos opciones para analizar una página; analizarla cuando se quiera desde el navegador y más tarde entrar a la aplicación de gestión y añadirla o analizarla empezando desde la aplicación, con lo que nos aparecerá el navegador y la aplicación esperará hasta que una página sea analizada.

Método 1	Analizar una página web desde el navegador	
FireEye	Desde el navegador nos situamos en la página y extraemos la información desde el panel del componente.	
Aplicación	Abrimos la aplicación y vamos a la ventana de Gestión de Sites. Seleccionamos la opción Añadir Site -> Usar última página analizada	
Aplicación	La aplicación nos muestra la ventana con la Ficha de la página analizada.	

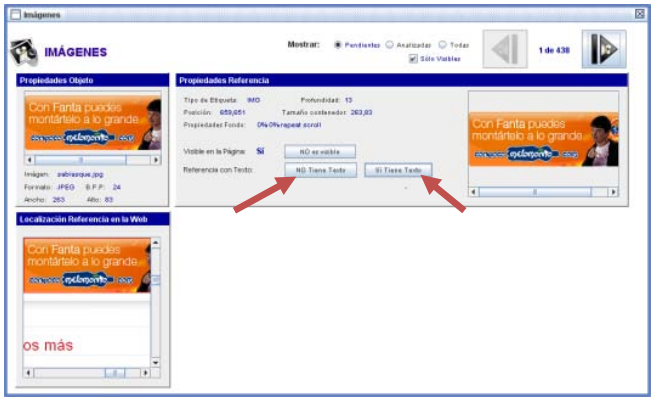
Método 2	Analizar una página web desde la aplicación
Aplicación	Abrimos la aplicación y vamos a la ventana de Gestión de Sites. Seleccionamos la opción Añadir Site -> Abrir FireFox  La aplicación abrirá el FireFox y esperará a que alguna página sea analizada por el componente FireEye.
FireEye	Desde el navegador nos situamos en la página y extraemos la información desde el panel del componente. 
Aplicación	La aplicación vuelve a aparecer automáticamente y nos muestra la ventana con la Ficha de la página analizada. 

- Clasificar imágenes

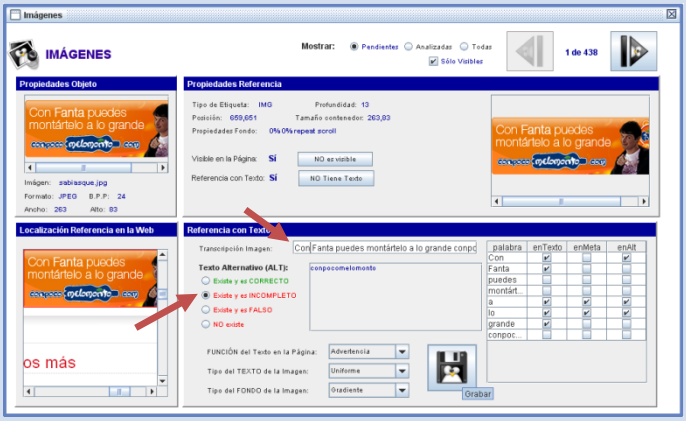
Una vez se ha analizado una página web deben clasificarse las imágenes. Este proceso se hace desde la ventana de imágenes. Su pueden ver las imágenes de una página o de todas según se haya accedido a ésta ventana desde la ficha de una página o desde la GUI principal respectivamente.

Clasificar imágenes.	
Desde la ficha de una web. Ir a ver/clasificar imágenes. 	Desde la pantalla principal Ir a imágenes. 

En esta pantalla se clasifican las imágenes de la página, hay que decir de cada imagen si tiene o no tiene texto.

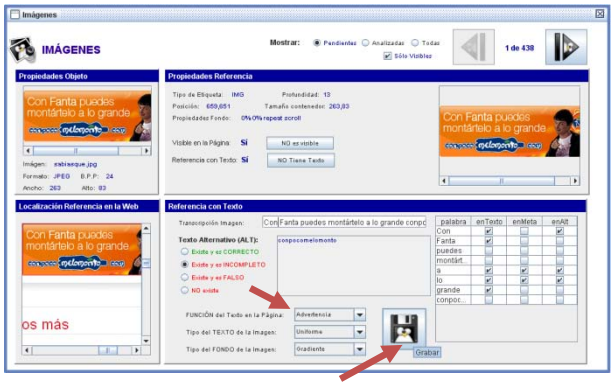


Si la imagen tiene texto habrá que escribir la transcripción de ese texto y seleccionar la opción que indica si se corresponde o no con el texto alternativo.



También se deberá seleccionar la función que realiza la imagen en la página, el tipo del texto de la imagen y el tipo de fondo.

Se graban los datos y aparecerá la siguiente imagen.
Cuándo no queden más referencias pendientes para clasificar, el estado de la página pasará automáticamente a Analizada.



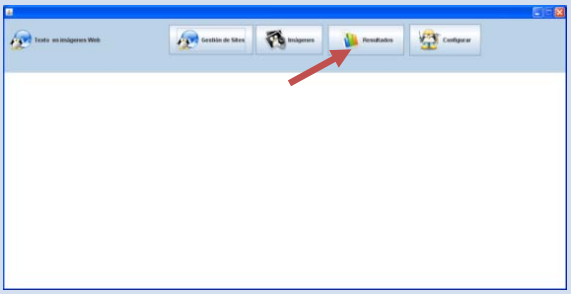
- Ver resultados

Con las páginas analizadas deben de verse los resultados para el estudio. Esto se hace en la ventana de Resultados. Aquí podemos crear selecciones de las páginas que se han analizado; se puede crear una selección con todas las página o por ejemplo una selección con sólo las páginas de una misma temática.

De cada selección podremos seleccionar los tipos de resultados que queremos ver

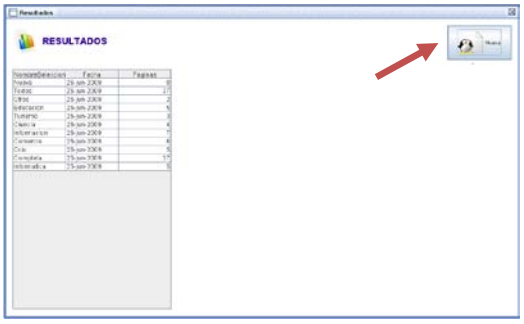
Ver resultados

Desde la pantalla principal Ir a resultados

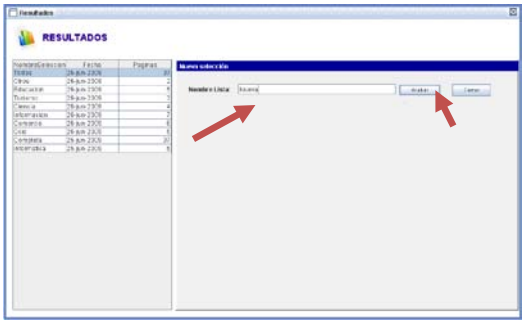


Para crear una nueva selección de páginas:

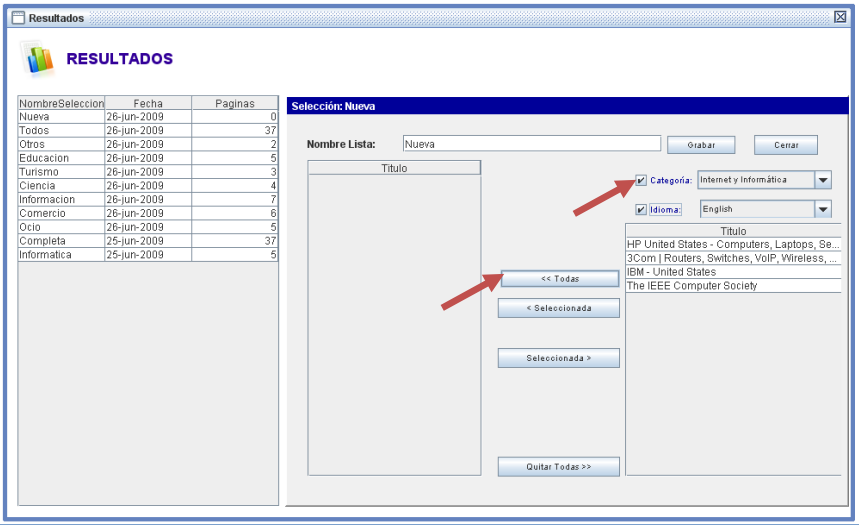
Paso 1: Seleccionar Nueva.



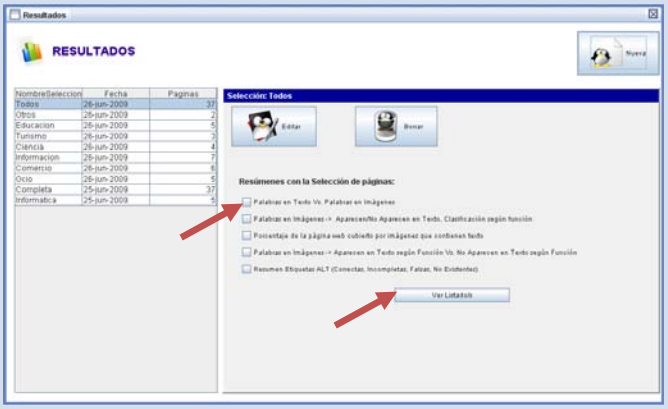
Paso 2: Poner el título y grabar

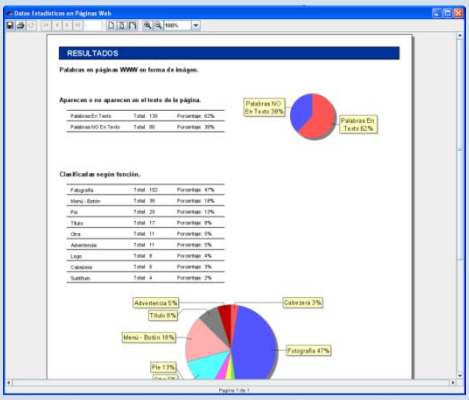


Paso 3: Seleccionar páginas, se pueden filtrar las páginas por temática y/o idioma.



Una vez con la selección creada, se escogen los resultados que se quieren visualizar.





5. Resultados

El software implementado nos ha permitido analizar páginas web, clasificar el contenido de las imágenes y sacar todos los resultados que nos interesan para nuestro estudio. A continuación presentaremos los resultados que hemos obtenido y los compararemos con otros estudios [1][2] anteriores.

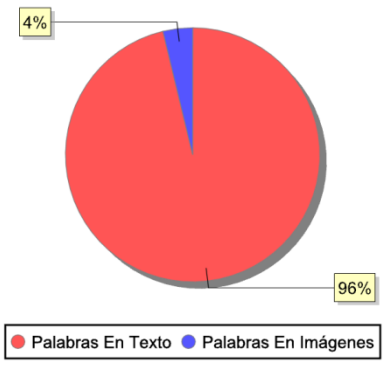
5.1 Situación actual de la web.

Para reflejar la situación actual de la web hemos analizado un conjunto de páginas de temáticas distintas. Deberemos asegurar que se cumplan los porcentajes de las clasificaciones de páginas web que habíamos realizado (apartado 3.2) tanto por temática de la página como por idioma.

Para tal efecto y no limitar cuantas páginas se pueden analizar de cada categoría, hemos optado en sacar unos primeros resultados para cada temática. Estos resultados los combinaremos al final, aplicando el peso correspondiente a cada temática, para obtener el resultado final que será representativo de la situación actual en Internet.
El conjunto de páginas analizadas se pueden ver en los anexos.
Las características que hemos analizado en el estudio y los resultados se exponen a continuación:

1. Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.

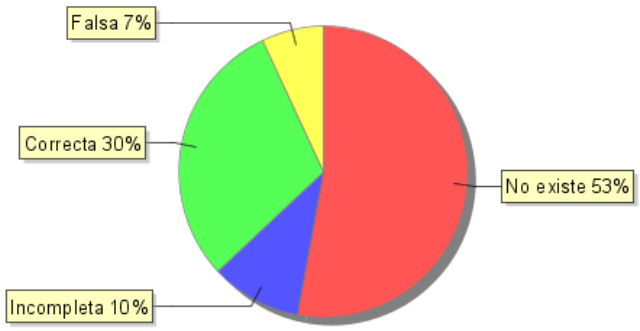
Categoría	Palabras en texto	Palabras en imágenes
Internet e Informática	99%	1%
Ocio	96%	4%
Comercio	92%	8%
Información	98%	2%
Ciencia y Tecnología	98%	2%
Turismo	95%	5%
Educación	94%	6%
Otros	93%	7%
Todas	**96%**	**4%**
Todas según pesos	**96%**	**4%**



Observamos que el porcentaje de palabras que aparece en imágenes (4%) es muy inferior al de palabras que aparecen en el texto de la página. Aún así, este porcentaje de palabras puede ser más significativo si estas palabras no aparecen en ningún otro sitio del texto y más aun si son importantes semánticamente para la página (logotipo, títulos..).

2. Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.

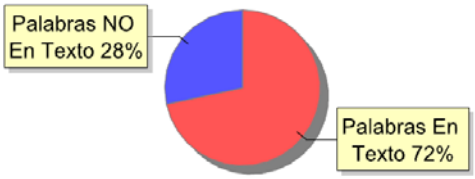
Categoría	Correctas	Incompletas	Falsas	No existentes
Internet e Informática	58%	17%	0%	25%
Ocio	13%	0%	3%	84%
Comercio	27%	17%	12%	44%
Información	44%	8%	7%	41%
Ciencia y Tecnología	25%	13%	8%	54%
Turismo	35%	15%	4%	46%
Educación	38%	0%	8%	54%
Otros	0%	10%	5%	85%
Todas	**30%**	**10%**	**7%**	**57%**
Todas según pesos	**28%**	**10,5%**	**5,5%**	**56%**



En este análisis vemos los peores resultados. De todas las imágenes que contienen texto, tan sólo el 28% de ellas tienen un texto descriptivo correcto. El 72% restante es falso (5,5%), incompleto (10,5%) o no existente (56%).

3. Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.

Categoría	Palabras en texto	Palabras NO en texto
Internet e Informática	63%	37%
Ocio	70%	30%
Comercio	72%	28%
Información	77%	23%
Ciencia y Tecnología	88%	12%
Turismo	71%	29%
Educación	69%	31%
Otros	67%	33%
Todas	**72%**	**28%**
Todas según pesos	**71%**	**29%**

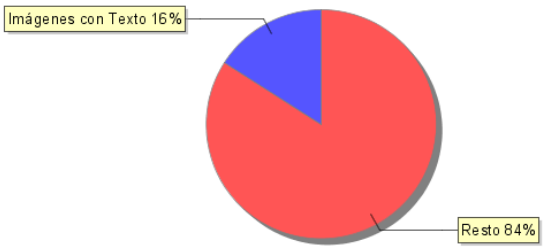


Casi un 30% de todas las palabras contenidas en imágenes no aparecen en ninguna otra parte del texto de la página, esto quiere decir que son totalmente inaccesibles y no podrán ser indexadas por los buscadores.

4. Porcentaje del área de la página cubierta por imágenes que contienen texto.

Categoría	Imágenes con Texto	Resto
Internet e Informática	34%	66%
Ocio	7%	93%
Comercio	37%	63%
Información	10%	90%
Ciencia y Tecnología	9%	91%
Turismo	7%	93%
Educación	16%	84%
Otros	12%	88%
Todas	**16%**	**84%**
Todas según pesos	**18%**	**82%**

Las imágenes que contienen texto cubren un area del 18% de la página web.



5.2 Comparación con un estudio [1] del año 2001

Vamos a comparar los resultados obtenidos en el estudio actual con otro estudio [1] realizado el año 2001.

- Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.

Categoría	Palabras en texto	Palabras en imágenes
Año 2001	83%	17%
Año 2009	96%	4%

Existe una diferencia bastante grande entre ambos resultados, pero esto es debido a que el primer estudio sólo contaba las palabras visibles en el texto de la página y el estudio actual cuenta tanto las palabras como las no visibles (texto oculto, texto de los meta tags..). Éste último resultado es más real porque todo este texto es el que indexan los buscadores.

- Análisis de las etiquetas descriptivas (ALT,TITLE).
Porcentaje de descripciones correctas e incorrectas.

Categoría	Correctas	Incompletas	Falsas	No existentes
Año 2001	44%	8%	3%	45%
Año 2009	28%	10,5%	5,5%	56%

Este resultado es uno de los más importantes para interpretar la evolución de este problema. Vemos que lejos de mejorar, ha empeorado; el 44% de imágenes con texto descriptivo correcto se ha reducido a un 28% y en cambio el porcentaje de imágenes sin texto descriptivo ha aumentado del 45% al 56%. Este aumento viene debido en gran medida a que entre un estudio y otro la web ha evolucionado, y ahora permite más opciones que antes para mostrar imágenes en una página web (CSS, JavaScript..). En estas nuevas opciones es más extraño que el autor de la web especifique texto descriptivo.

Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.

Categoría	Palabras en texto	Palabras NO en texto
Año 2001	24%	76%
Año 2009	71%	29%

Al igual que en la primera característica comparada vemos que hay una gran diferencia entre un estudio y el otro. Pero esto es debido precisamente a lo que ya comentamos anteriormente de que en el estudio actual tenemos en cuenta más palabras que sólo las visibles de la página.

6. Conclusiones

Hemos conseguido realizar los objetivos de este proyecto en el plazo de tiempo estimado en la planificación del mismo. A continuación repasaremos cuales eran estas objetivos y terminaremos comentando el trabajo futuro que se podría realizar.

6.1 Objetivos logrados

Los objetivos específicos que teníamos para este proyecto eran:

O1. Escoger un conjunto de páginas que sean representativas del estado actual en la Web.
Logrado. Se analizaron varios estudios sobre Internet y se realizó una clasificación por temáticas de páginas e idiomas que es representativa del estado actual de Internet. (apartado 3.2)

O2. Diseñar una base de datos para almacenar toda la información significante de las páginas web analizadas.
Logrado. Se ha diseñado una base de Datos en MySQL que almacena toda la información de las páginas, así como todas sus imágenes y una captura de la misma.

O3. Implementar un software que ayude, guiando al usuario, al análisis y a la clasificación de las páginas y a las imágenes de éstas.
Logrado. Hemos desarrollado un software que responde a estas funcionalidades y con una interface muy amigable.

O3.1. Extracción de Información de manera automática de una página web, como texto e imágenes y propiedades de estas (tipo, dimensiones, bpp).
Logrado. Tal como se ha comentado en el apartado 4.3, hemos desarrollado una extensión para el navegador que se encarga de extraer toda la información que nos interesa de una página web.

O3.2. Introducción manual de Información. Deben poder clasificarse las imágenes según si tienen texto o no y en el caso de que tengan texto habrá que poner la transcripción de ése texto.
Logrado. El software desarrollado tiene una interface que permite y facilita mucho la clasificación de las imágenes.

O3.3. Sacar resultados de las páginas analizadas. Permitir crear varias selecciones de páginas y con ellas analizar las características que nos interesan.
Logrado. El software tiene un apartado dedicado a mostrar los resultados de cualquier selección de páginas que se quiera. Estos resultados se pueden ver gráficamente e imprimir o exportar a PDF.

O4. Analizar la situación actual de la web. Mediante el software desarrollado deberá analizarse el conjunto de páginas representativas del estado actual en la Web y sacar los siguientes resultados:

O4.1. Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.

O4.2. Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.

O4.3. Porcentaje de palabras contenidas en imágenes que no aparecen en el texto de la página.

O4.4. Porcentaje del área de la página cubierta por imágenes que contienen texto.

Logrado. Todos estos resultados han sido analizados (apartado 5.1).

O4.5. Comparar los resultados con un estudio [1] del año 2001.

Logrado. La comparación con un estudio [1] del año 2001 se ha realizado y se han comentado las conclusiones (apartado 5.2).

6.2 Trabajo futuro.

En el proceso de desarrollo del software ya se tuvieron en cuenta algunas aplicaciones futuras. Por eso, el software extrae más información de una página web de la que sería necesaria para el estudio que hemos realizado.

Una de esas posibles aplicaciones futuras sería la de saber qué están mirando los usuarios en un página web. Este estudio sería muy importante para creadores de páginas web, diseñadores de interfaces, anunciantes, etc.. Para ello se necesitaría de un dispositivo EyeTracker. Con la información que obtenemos de nuestra aplicación y los datos que nos proporciona el EyeTracker podríamos tanto ver un heat-map de lo que el usuario ha visto en la página, como de crear clasificaciones de qué imágenes ha visto más y durante cuánto tiempo.

Otra aplicación que también se ayudaría de la información extra que obtiene nuestra aplicación es la de reconocimiento de texto en imágenes; para que esta técnica sea eficiente necesita de información Ground Truth. La base de datos que hemos creado contiene muchas imágenes con texto e información sobre ellas como: propiedades básicas (dimensiones,tipo,bpp), tipo del texto (uniforme, multicolor..), tipo de fondo (uniforme,gradiente, fotográfico..) y la transcripción del texto. Ésta transcripción se podrá usar para saber si el reconocimiento de texto que se haga de la imagen es correcto o no.

Otro aspecto en que se podría trabajar es en ampliar el software desarrollado para que tenga en cuenta más objetos de los que aparecen en una página web que sólo las imágenes. El contenido multimedia ha aumentado mucho en la web y éstos contenidos presentan los mismos problemas de indexación / inaccesibilidad de las imágenes. Si se tuvieran en cuenta estos contenidos y se volviera a realizar el estudio de este proyecto veríamos que el problema aún es mayor.

7. Referencias

[1] .Karatzas - A.Antonacopoulos - J.Ortiz Lopez: “Accessing Textual Information Embedded in Internet Images” *Proceedings of SPIE, Internet Imaging II, San Jose, USA, January 2001, Vol. 4311, pp. 198-205*

[2] T. Kanungo, C. H. Lee and R. Bradford, "What Fraction of Images on the Web Contain Text?", Proc. of Int. Workshop on Web Document Analysis, Seattle,WA, Sept. 8, 2001.

[3] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson and Gordon L. Hempton: “WebInSight: Making Web Images Accessible” Department of Computer Science and Engineering University of Washington Seattle

[4] D. Lopresti and J. Zhou. Document analysis and the World WideWeb. In Proceedings of IAPR Workshop on Document Analysis Systems, Marven, PA, 1996.

[5] J. Zhou and D. Lopresti. Extracting text from WWW images. In Proceedingsof the IAPR Internation Conference on Document Analysis Recognition, Ulm, Germany, 1997.

[6] D. Lopresti and J. Zhou. Locating and recognizing text in WWW images. Information Retrieval, 2:177–206, 2000.

[7] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In Proceedings of the 2nd ACM International Conference on Digital Libraries, 1997.

[8] H. Li, O. Kia, and D. Doermann. Text enhancement in digital video. In Proceedings of SPIE Conference on Document Recognition IV, pages 1–8, 1999.

[9] R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In Proceedingsof ACM Multimedia, pages 11–20, 1996.

[10] www.w3c.org/wai/ W3C: Web Accessibility Initiative, 2006.

[11] www.fbbva.es/TLFU/dat/Estudio_Internet_2008.pdf Estudio de la Fundación BBVA sobre el estado de Internet en España Mayo 2008

[12] www.hispanom.com Diciembre 2008

[13] www.internetworldstats.com enero 2008

8. Anexos

Todos los anexos están disponibles en formato PDF en el CD adjunto a la memoria.

Anexo 1. Listado de páginas web analizadas

Página	URL	Categoría	Idioma	Numero Imágenes	Numero Palabras Texto	Numero Palabras Imagenes
The University of Liverpool	http://www.liv.ac.uk/	Educación	English	2	162	6
HP United States - Computers, Laptops, Servers, Printers and more	http://www.hp.com/#Product	Internet y Informática	English	2	606	26
3Com | Routers, Switches, VoIP, Wireless, Network Management	http://www.3com.com/	Internet y Informática	English	5	387	40
IBM - United States	http://www.ibm.com/us/en/	Internet y Informática	English	2	853	13
The University of Manchester	http://www.manchester.ac.uk/	Educación	English	2	201	4
Laptop, Notebook, Desktop, Server and Embedded Processor Technology	http://www.intel.com/#/es_ES_01	Internet y Informática	Castellano	2	213	2
The New York Times - Breaking News, World News & Multimedia	http://www.tlmes.com/	Información	English	10	2526	21
#GD' HJGD' (CE AJ EHB9 ,1#) 9EF	http://www.omandaily.com/	Información	Otro	6	372	9
WWF - WWF España	http://www.wwf.es/	Información	English	8	614	35
Coca-Cola: The Coca-Cola Company	http://www.thecoca-colacompany.com/	Comercio	English	10	137	38
Coca-Cola - happing - La forma más divertida de conocer gente	http://www.cocacola.es/	Comercio	Castellano	7	329	42
Flights, Hotels and Holidays with British Airways - BA.com	http://www.britishairways.com/travel/home/public/en_gb	Comercio	English	9	803	56
Flight booking, hotels and car hire - easyJet.com	http://www.easyjet.com/en/book/index.asp	Comercio	English	12	1009	67
Lufthansa - Book your flight online | Flights to Europe from £49 | Fly wor	http://www.lufthansa.com/online/portal/lh/uk	Comercio	English	2	380	9
BBC - Homepage	http://www.bbc.co.uk/	Información	English	2	275	4
Yahoo! UK & Ireland Eurosport - Sports News | Live Scores | Sport	http://uk.eurosport.yahoo.com/	Información	English	10	1340	22
Channel 4	http://www.channel4.com/	Ocio	English	2	265	1
The IEEE Computer Society	http://www2.computer.org/portal/web/guest/home	Internet y Informática	English	1	3812	2
The Official Home Page for All Things Disney | Home | Disney.com	http://home.disney.go.com/	Ocio	English	1	23	1
uefa.com	http://www.uefa.com/	Información	English	26	1789	64
JCPenney	http://www.jcpenney.com/jcp/default.aspx	Comercio	English	19	435	88
UCL - London's Global University	http://www.ucl.ac.uk/	Educación	English	2	234	4
Homepage - University of Oxford	http://www.ox.ac.uk/	Educación	English	12	218	16
Hotels | London.com	http://www.london.com/es	Turismo	Castellano	4	1022	14
Manchester UK Guide	http://www.manchester.com/	Turismo	English	18	633	103
Vodafone: Telefonía Móvil	http://www.vodafone.es/particulares/	Ciencia y Tecnología	Castellano	6	389	41
Landis+Gyr | L+G North America	http://www.landisgyr.com/na//en/pub/index.cfm	Ciencia y Tecnología	English	1	212	6
Universitat Autònoma de Barcelona	http://www.uab.es/	Educación	Otro	6	344	59
Apuestas Deportivas, Casino, Poker y Juegos Mlapuesta.com	http://www.mlapuesta.com/	Otros	Castellano	20	595	98
Google	http://www.google.es/	Información	Castellano	1	64	0
Abbey Road Studios - Abbey Road Studios	http://www.abbeyroad.co.uk/	Ciencia y Tecnología	English	14	204	22
Ferrari - Il sito Ufficiale della Casa Automobilistica di Maranello	http://www.ferrari.com/Italian/Pages/Home.aspx	Ocio	Otro	3	378	2
Honda Cars Motorcycles Watercraft ATVs Engines Generators, Acura	http://www.honda.com/	Ocio	English	2	291	14
Home - ABC.com	http://abc.go.com/	Ocio	English	29	276	47
NASA - Home	http://www.nasa.gov/	Ciencia y Tecnología	English	3	2322	3
Edinburgh.com - EDINBURGH.com: The official dot com for EDINBURGH	http://www.edinburgh.com/	Turismo	English	4	808	9
YouTube - Broadcast Yourself.	http://www.youtube.com/	Otros	Castellano	1	507	4

Anexo 2. Resultados de nuestro estudio - Porcentaje de palabras en páginas WWW contenidas en imágenes o en el texto.

RESULTADOS

Palabras en páginas WWW en forma de imágen y en todo el Texto.

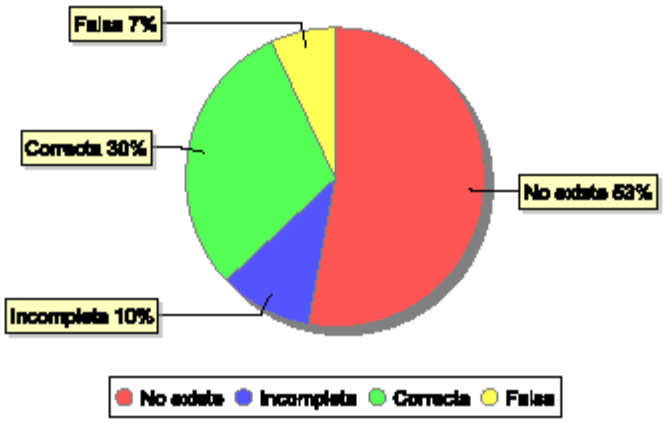
Palabras En Texto	Total:	28018	Porcentaje:	97%
Palabras En Imágenes	Total:	992	Porcentaje:	3%

Anexo 3. Resultados de nuestro estudio - Análisis de las etiquetas descriptivas (ALT,TITLE). Porcentaje de descripciones correctas e incorrectas.

RESULTADOS

Resumen Etiquetas ALT (Correctas, Incompletas, Falsas, No Existentes)

No existe	Total:	136	Porcentaje:	53%
Correcta	Total:	77	Porcentaje:	30%
Incompleta	Total:	26	Porcentaje:	10%
Falsa	Total:	18	Porcentaje:	7%



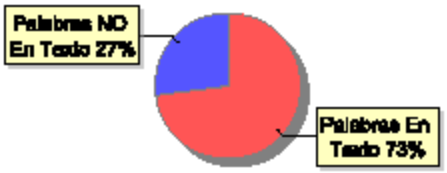
Anexo 4. Resultados de nuestro estudio - Análisis de las palabras en forma de imagen que aparecen o no aparecen en el texto de la página.

RESULTADOS

Palabras en páginas WWW en forma de imágen.

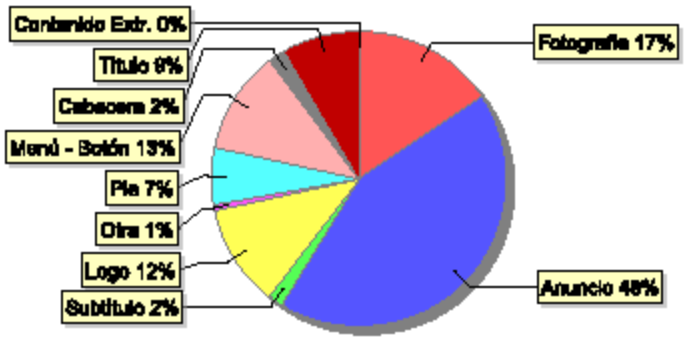
Aparecen o no aparecen en el texto de la página.

Palabras En Texto	Total: 793	Porcentaje: 73%
Palabras NO En Texto	Total: 294	Porcentaje: 27%



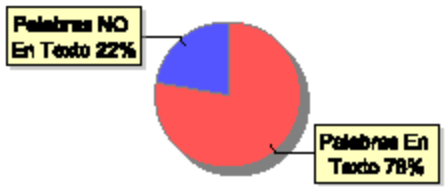
Clasificadas según función.

Anuncio	Total: 476	Porcentaje: 48%
Fotografía	Total: 173	Porcentaje: 17%
Menú - Botón	Total: 125	Porcentaje: 13%
Logo	Total: 123	Porcentaje: 12%
Título	Total: 93	Porcentaje: 9%
Pie	Total: 69	Porcentaje: 7%
Cabecera	Total: 20	Porcentaje: 2%
Subtítulo	Total: 18	Porcentaje: 2%
Otra	Total: 7	Porcentaje: 1%
Contenido Extr.	Total: 0	Porcentaje: 0%

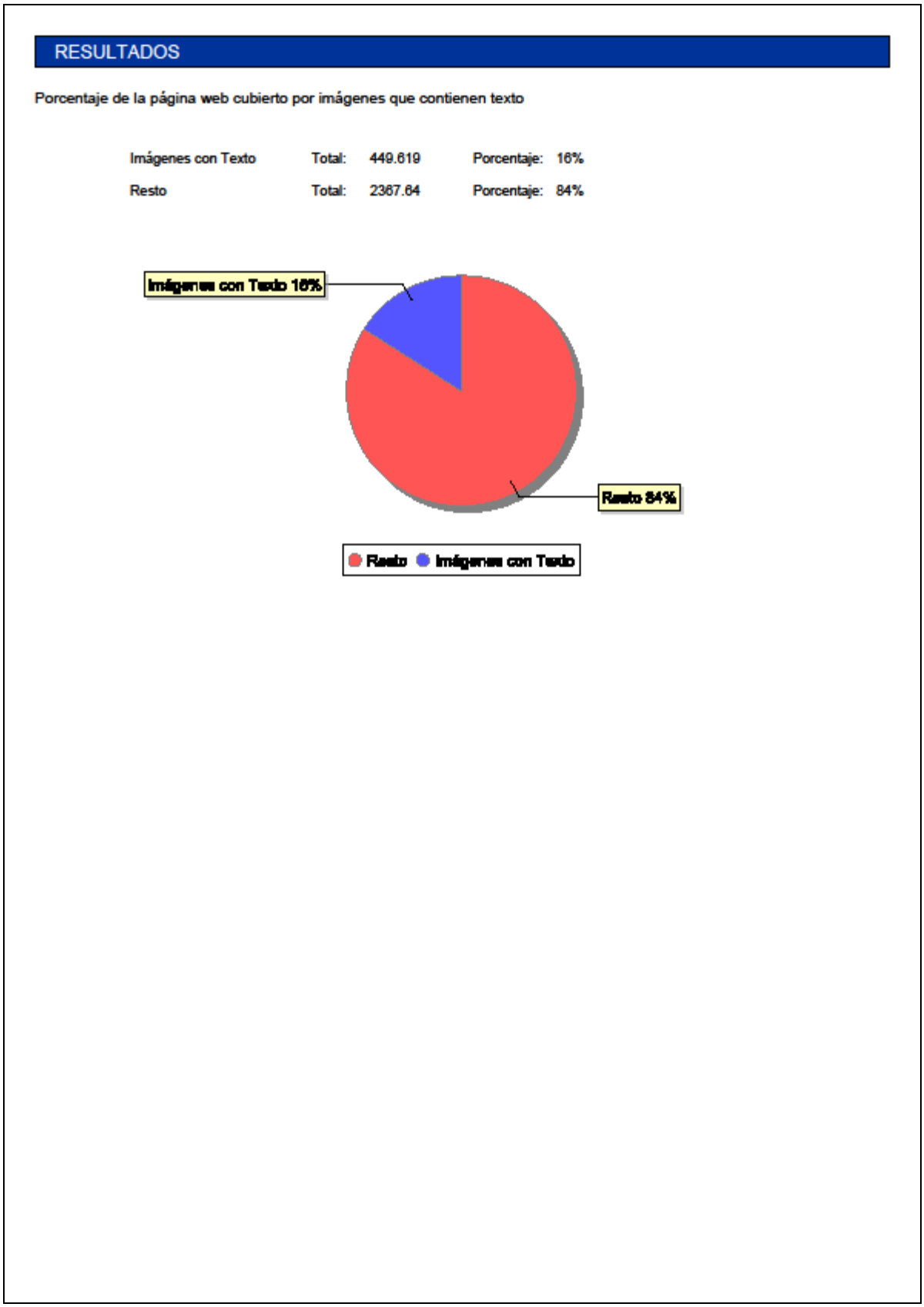


Aparecen o no en el texto de la página según el peso de la función.

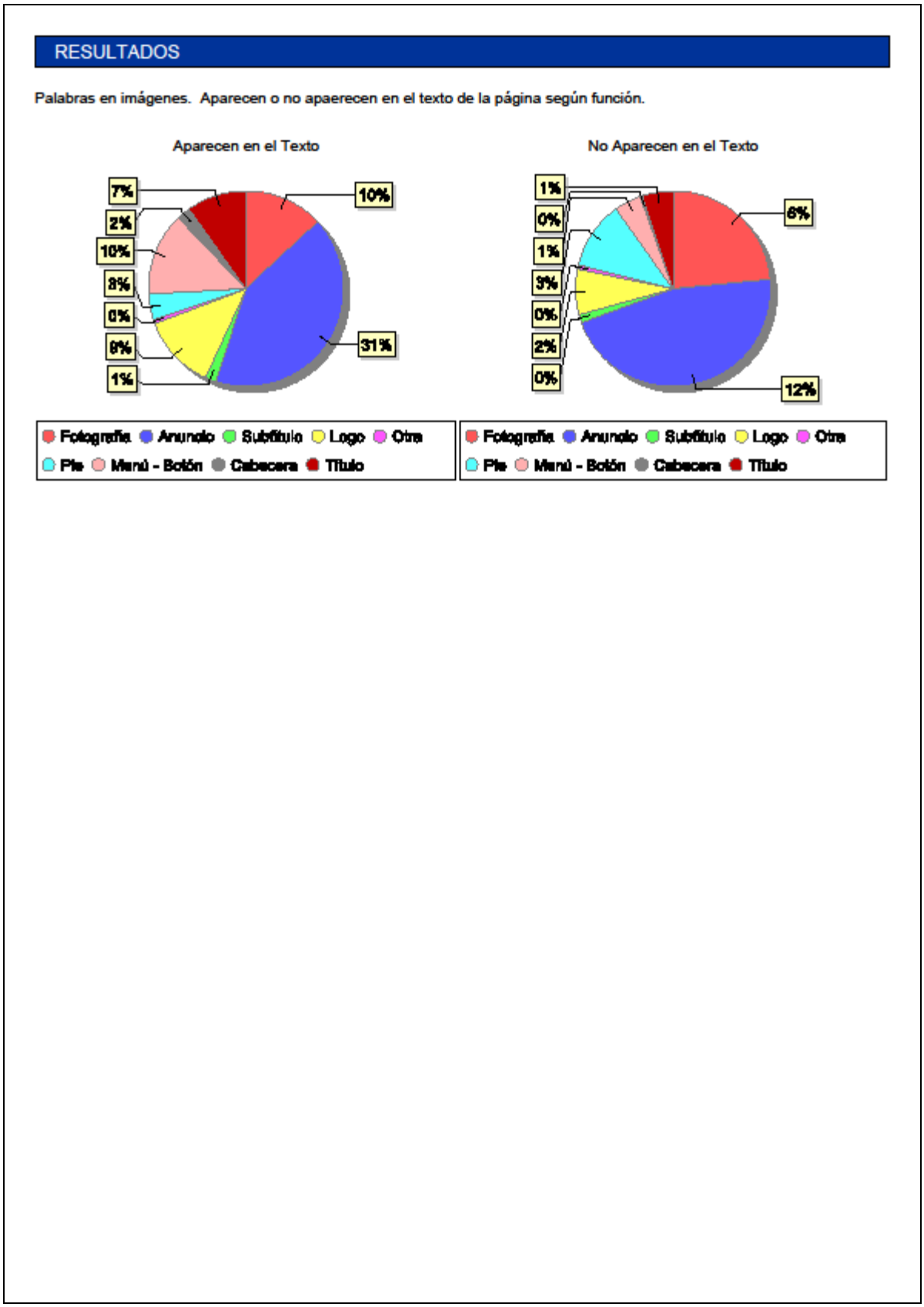
Palabras En Texto	Total: 10175	Porcentaje: 78%
Palabras NO En Texto	Total: 2910	Porcentaje: 22%



Anexo 5. Resultados de nuestro estudio - Porcentaje del área de la página cubierta por imágenes que contienen texto.



Anexo 6. Resultados de nuestro estudio – Palabras contenidas en imágenes que aparecen/No aparecen en el texto de la página según la función que realiza la imagen.



Resum:

L' indexació i la búsqueda de pàgines web es basa en l'anàlisi de text. La tecnologia actual encara no pot processar d'una maner eficient i suficientment ràpida el text contingut en les imatges de les pàgines web. Aquest fet planteja un problema important d'indexació però també d'inaccessibilitat.

Per poder quantificar aquest problema hem desenvolupat una aplicació software que ens permet realitzar un estudi sobre aquesta situació. Hem utilitzat aquest software per analitzar un conjunt de pàgines web representatives de la situació actual a Internet. Aquests resultats obtinguts s'han analitzat i comparat amb estudis anteriors.

Resumen:

La indexación y la búsqueda de páginas web se basan en el análisis de texto. La tecnología actual, aún no puede procesar de una manera eficiente y suficientemente rápida el texto contenido en las imágenes de las páginas WWW. Este hecho plantea un problema importante de indexación pero también de inaccesibilidad.

Para poder cuantificar este problema hemos desarrollado una aplicación software que nos permite realizar un estudio sobre esta situación. Hemos utilizado este software para analizar un conjunto de páginas web representativas de la situación actual en Internet. Estos resultados obtenidos se han analizado y comparado con estudios anteriores.

Summary:

Indexing and searching for WWW pages is relying on analyzing text. Current technology cannot process in an efficient way and quickly enough the text embedded in images on WWW pages. This fact is a significant indexing problem but inaccessibility too.

To quantify this problem we have developed a software application that allows us to conduct a study on this. We have used this software to analyze a set of web pages representing the current Internet situation. These results have been analyzed and compared with previous studies.