UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

SISTEMA DE SEGUIMIENTO Y ANÁLISIS DE MEDIOS DE COMUNICACIÓN EN INTERNET

PROYECTO FIN DE CARRERA

Ingeniería Técnica de Telecomunicación

Sonido e Imagen

Autor: Daniel López Fuentes

Tutor: Julio Villena Román

Octubre 2010

Título: Sistema de seguimiento y análisis de medios de comunicación en Internet.

Autor: Daniel López Fuentes

Tutor: Julio Villena Román

EL TRIBUNAL

Presidente: José Alberto Hernández Gutiérrez

Secretario: Iria Estévez Ayres

Vocal: Soledad Escolar Díaz

Realizado el acto de defensa del Proyecto Fin de Carrera el día 28 de Octubre de 2010 en Leganés, en la escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de:

Fdo.: Presidente Fdo.: Secretario Fdo.: Vocal

AGRADECIMIENTOS

En primer lugar dar las gracias a mi tutor, Julio Villena, por toda su dedicación y tiempo invertido en este proyecto, sin los cuales nunca podría haber visto la luz.

A mis padres darles las gracias por todo, absolutamente por todo, ya que han logrado convertirme en la persona que soy, y además por ser los mejores mecenas que un hijo podría desear. Este proyecto es más vuestro que mío, otra vez gracias.

A Paloma, porque hasta el consejo más pequeño siempre fue un gran consejo, este proyecto también tiene una parte tuya, gracias, guapa.

A toda la gente que ha conseguido que estos 6 años de universidad se conviertan en el sueño de una noche. Gracias Tamara, por tu amistad incondicional y tu sabiduría invertida en mí. Gracias Lola, porque solo el hecho de tenerte cerca ya es un regalo. Gracias Irene, por tu ayuda y tu risa, sin ellas yo no sería el mismo. Gracias Nerea, siempre serás la llave de la puerta que tanto miedo me daba abrir. Gracias a Lorena y JC, Teje, Clara, Lucía, Sarita, Pato, Mónica, Ana B., Carlos, Laurita, Marian, Rocío (Beca), Tefa, y muchos más de los que alguna vez tuve su apoyo, gracias por todas esas interminables horas de prácticas compartidas y por todas esas increíbles horas de fiesta.

No puedo olvidarme de ese rayo de luz, en forma de 4 chicas, que llegó a mi vida hará un año y sin las cuales seguiría perdido en un mar de dudas. Gracias Anita (Osezno) porque irradias tanta belleza que hasta un simple mortal a tu lado es un dios. Gracias Olaia, por ser el mejor diario que un amigo podría desear, sabes que cada lagrima y risa a tu lado para mí es un día de vida. Y gracias Meri y Ana, por hacerme sentir tan cómodo a vuestro lado, nunca imaginé tener unas amigas tan increíbles.

Y por último pero no menos importante agradecer a mis amigos de toda la vida, sus risas, aventuras, sus secretos… Gracias Elena, una frase no puede mostrar todo lo que siento por ti, así que solo decirte gracias. Gracias Adri, porque la sangre no es la única razón para llamarse hermano. Gracias Rafa, la definición de amigo debería ser tu nombre. Gracias Nora, por hacerme sonreír constantemente, eres un amor. Gracias Amanda, por demostrarme que los mejores amigos están a la vuelta de la esquina. Gracias Carmen, nunca pensé que una rubia pudiera volverme el corazón loco. Y gracias a todos los que con vuestros consejos habéis conseguido crear la persona que soy y por consiguiente mis logros como es este proyecto.

RESUMEN

La búsqueda constante del hombre por satisfacer cada vez mejor su necesidad de comunicación ha sido el impulso que ha logrado la instauración en el mundo de instrumentos cada día más poderosos y veloces en el proceso comunicativo. Solo basta una mirada al pasado para ver cómo el ser humano ha logrado evolucionar sus formas de comunicación, desde los rudimentarios métodos como la escritura jeroglífica hasta la aparición del teléfono, el cine, la radio o la televisión.

En los últimos años, y debido al aumento exponencial de páginas web, se ha convertido en una odisea recuperar y organizar la gran cantidad de información existente en Internet. Así nace lo que hoy se conoce como buscadores, sistemas informáticos que indexan archivos almacenados en servidores web, gracias a programas araña encargados de inspeccionar páginas del World Wide Web de forma metódica y automática, y que proporcionan una manera de obtener la información deseada de forma eficiente y sencilla.

El Proyecto Fin de Carrera desarrollado consiste en el diseño e implementación de un sistema web de seguimiento de noticias en medios de comunicación en español en Internet, a partir de la descarga periódica de sus canales RSS.

El sistema de seguimiento de noticias consta de dos módulos. En primer lugar, la creación de un robot encargado de visitar una lista de RSS, recolectar la información de las páginas de cada noticia, y finalmente almacenar dicha información en una base de datos. Y en segundo lugar un motor de búsqueda o buscador web que se encarga de buscar las palabras clave que el usuario indica en la búsqueda, en los datos almacenados en la base de datos y mostrarlos al usuario de manera comprensible y organizada.

ÍNDICE DE CONTENIDOS

1. Introducción 1

1.1 - Motivación 1
1.2 - Objetivos 3
1.3 - Estructura de la memoria 3

2. Estado del arte 5

2.1 - Introducción 5
2.2 - Recuperación de Información 8
2.3 - Modelos de Recuperación 10
2.4 - La Recuperación de Información en Internet 13
2.5 - Tipos de Buscadores 15
2.5.1 - Directorios 17
2.5.2 - Motores de búsqueda 19
2.6 - Los Buscadores del Futuro 27
2.7 - Medios de Comunicación en Internet 29
2.8 - La Web 2.0 32
2.9 - Sistemas Web de Seguimiento y Análisis de Medios de Comunicación en el Mercado 35

3. Tecnologías Involucradas 37

3.1 - Introducción 37
3.2 - Html, Xml, Xhtml 37
3.3 - CSS 40
3.4 - Páginas Web Dinámicas 40
3.5 - Lenguajes de Programación 41
3.5.1 - JavaScript 41
3.5.2 - CGI 41
3.5.3 - PHP 42
3.5.4 - Asp.Net 43
3.5.5 - JSP y Servlets 44
3.5.6 - Elección de PHP frente a otros Lenguajes 45
3.6 - Bases de Datos 46
3.6.2 - SQL 47
3.6.3 - MySQL 47
3.6.4 - PostgreSQL 48
3.6.5 - JDBC 49
3.6.6 - Elección de MySQL como Sistema de Administración de Bases de Datos 50

3.7 - Recuperación de Información: Sistemas Basados en Texto 51
3.7.1 - Swish-E 51
3.7.2 - Xapian 52
3.7.3 - Lucene 53

4. Desarrollo del Proyecto 56
4.1 - Arquitectura del Sistema 56
4.2 - Implementación del Sistema 61
4.2.1 - Araña 61
4.2.2 - Base de datos 66
4.2.3 - Indexador 69
4.2.4 - Interfaz de recuperación web 69
4.2.5 - Función de análisis de noticias 76

5. Validación del sistema 79
5.1 - Pruebas de indexación 79
5.2 - Pruebas de búsqueda 80

6. Presupuesto 83

7. Conclusiones y Trabajos Futuros 84
5.1 - Conclusiones 84
5.2 - Trabajos Futuros 86

Bibliografía y Referencias 88

ÍNDICE DE IMÁGENES Y TABLAS

Figura 1: Pirámide de la Información 8
Figura 2: Clasificación de los Modelos de Recuperación..... 10
Figura 3: Representación en forma de árbol de la consulta (t1 y t3) o (t5 y NO(t7)) 11
Figura 4.1: Modelos listas no sobrepuestas 13
Figura 4.2: Modelos nodos proximales 13
Figura 5: Estructura de un motor de búsqueda 20
Figura 6: Comparativa entre "Google" y "Yahoo" Keyword: "uc3m"..... 23
Figura 7: Web 1.0 vs Web 2.0..... 32
Figura 8: Ejemplo de código RSS 34
Figura 9: News Brief..... 35
Figura 10: Tus Titulares..... 36
Figura 11: Petición de una página en PHP..... 42
Figura 12: Logo Swish-e 52
Figura 13: Logo Xapian 52
Figura 14: Logo Lucene 54
Figura 15: Arquitectura del motor de búsqueda..... 57
Figura 16: Presentación de la información en RSS 61
Figura 17: Código con la cabecera de información en cada RSS 62
Figura 18: Información en cada noticia del RSS 63
Figura 19: Código SimplePie 65
Figura 20: Ejemplo de código Cron..... 65
Figura 21: Código Acceso a BD 66
Figura 22: Uso de la sentencia SELECT 69
Figura 23: Ejemplo de regla CSS 70
Figura 24: Vista Principal: Noticias Diarias 70
Figura 25: Calendario..... 71
Figura 26: Resultado de la búsqueda para el término: España (18/11/09 - 4/02/10)... 73
Figura 27: Gráfico del resultado de la búsqueda del término: España 76
Figura 28: Archivos incluidos en el fichero grafico.php 77
Figura 29: Código para la creación de la función grafico.php 78
Figura 30: Detalle de prueba Nº3 81

Tabla 1: Relación: Noticias..... 68

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

El texto es la principal forma de almacenamiento del conocimiento humano y, después del lenguaje hablado, también es la principal forma de divulgarlo y transmitirlo. Las técnicas para almacenar y buscar los documentos escritos son tan antiguas como el propio lenguaje escrito. Desde hace, aproximadamente, cuatro mil años, el hombre ha organizado la información para después recuperarla y usarla con facilidad.

Dado que la información es estratégica hoy en día, ya sea a nivel de empresa, geográfico o incluso de estado, se ha convertido en aspectos indispensables en almacenamiento, acceso y explotación.

En concreto, los artículos de prensa constituyen gran parte de la información a la que usuarios de todo el mundo acceden cada día.

Si a esto se añade que la red, hoy en día constituye el principal medio a través del cual se puede tener acceso a grandes volúmenes de extensa y diversa información, y que cualquier persona o entidad puede ser tanto usuaria como productora, se ve la necesidad de crear nuevos mecanismos para acceder posteriormente a los documentos almacenados.

En este contexto nacen los buscadores, que intentan solucionar el problema que el navegante de Internet tiene, aquél que siempre ha acompañado al ser humano, independientemente de la fase tecnológica en la que haya vivido. En cualquier momento, y para cualquier acción o decisión, tiene que buscar, localizar y recuperar la información que le permita desarrollar sus actividades. Así herramientas como la mundialmente famosa Google han prosperado por esta causa.

A través de las distintas técnicas de búsqueda desarrolladas, los usuarios de Internet pueden acceder de forma rápida y sencilla a todo tipo de información generada en cualquier lugar del mundo. Tan sólo hay que saber cómo buscar. Y es que muchas de las personas que se aventuran en el apasionante mundo de Internet intentando encontrar un dato concreto entre los más de seiscientos mil millones de sitios web [24] disponibles en este abismo de información, se sienten frustradas al comprobar que el resultado de su búsqueda en buscadores populares son demasiado generales. No es un problema de falta de información, sino precisamente de lo contrario. La sobrecarga de datos que inunda la red provoca muchas veces el efecto de la desinformación, por muy paradójico que pueda parecer. Sin embargo, éste es un problema con solución.

Es interesante que el usuario conozca que existen más recursos de búsqueda, además de los habituales, sobre todo si se enfrenta a búsquedas concretas. Y es aquí donde los buscadores especializados, se han convertido en la herramienta más eficaz para afrontar con éxito estas pesquisas y de donde precisamente nace el motivo que guía la creación del proyecto: un sistema de seguimiento y análisis de medios de comunicación que permita mantener al usuario informado de la actualidad española de una manera rápida y concisa.

A la hora de elegir un tema concreto para la creación del buscador especializado se optó por escoger las noticias publicadas por distintos medios de comunicación en Internet, ya que representan una gran fuente de información en constante crecimiento y además porque en la actualidad Internet está pasando de ser simplemente una herramienta funcional a convertirse en un medio principal, donde un impresionante 83% de los internautas considera que es mejor leer una noticia en Internet que en los periódicos impresos[31].

Si bien en el mercado existen soluciones parecidas al diseño creado, el valor añadido a la información que se muestra es un seguimiento de los medios, categorizando la información y recopilándola de tal forma que pueda ser analizada fácilmente.

1.2 OBJETIVOS

El objetivo fundamental de este Proyecto Fin de Carrera es desarrollar un sistema de seguimiento y análisis de medios de comunicación en internet mediante la implementación de un buscador de noticias, basado en una lista de RSS, y la creación de una herramienta de análisis que muestre la evolución de las noticias en el tiempo.

En este sentido, esta aplicación se plantea como una herramienta de apoyo, que permita manejar al usuario grandes volúmenes de información de forma rápida y eficaz, así como poder procesarla fácilmente gracias al seguimiento de medios realizado.

Mediante el empleo de técnicas de indexación, el sistema podrá satisfacer las consultas efectuadas en el buscador, las cuales consistirán en la introducción de una o varias palabras clave y un rango de fechas donde enmarcar la búsqueda. Asimismo, el sistema debe devolver, ante dichas consultas de usuario, aquellas noticias que concuerden con respecto a la solicitud de información.

Adicionalmente como objetivos secundarios, se requerirá que el sistema sea eficiente en la resolución de búsquedas, y la eficiencia del resto de tareas ejecutadas previamente se evaluará también de acuerdo a este rendimiento final. Además, debe existir un compromiso con la agilidad y rapidez como parámetros importantes de calidad en la aplicación.

1.3 ESTRUCTURA DE LA MEMORIA

La memoria de este proyecto se ha dividido en cinco capítulos, a continuación se incluye una breve explicación del contenido de cada uno de ellos, con el objetivo de ubicar al lector en las distintas partes que la componen.

- **CAPÍTULO 1:** tras una explicación sobre los fundamentos y motivaciones presentes en un sistema de seguimiento y análisis de medios de comunicación en Internet, se

enumeran los objetivos perseguidos en este Proyecto Fin de Carrera. Y por último, se describe la estructura de la memoria realizada.

- **CAPÍTULO 2:** en este capítulo se enmarca el contexto sobre el que se crea el proyecto, así como una visión general de todo sistema de recuperación de información. Además se explican los tipos de buscadores que existen en la actualidad y la evolución que están experimentando con el crecimiento de la red. Y finalmente se hace un repaso de las herramientas que poseen hoy en día los medios de comunicación en Internet para distribuir información
- **CAPÍTULO 3:** dado la gran cantidad de tecnologías y lenguajes de programación existentes en la actualidad para el desarrollo de aplicaciones web, es conveniente la creación de este capítulo donde se da una visión general de estas tecnologías, así como una explicación que justifique su elección en la creación del proyecto.
- **CAPÍTULO 4:** en este cuarto capítulo se describe la arquitectura del sistema desarrollado, explicando cada una de las fases que lo compone: desde la fase de extracción de datos, pasando por la creación de la base de datos, hasta la posterior representación de la información y su análisis temporal.
- **CAPÍTULO 5:** validación del sistema, recoge la descripción de las pruebas efectuadas para verificar de forma empírica el funcionamiento del sistema, de acuerdo a los parámetros previstos.
- **CAPÍTULO 6:** presupuesto del proyecto.
- **CAPÍTULO 7:** finalmente, se explican las conclusiones a las que se ha llegado durante la realización de este proyecto. Además se presentan posibles líneas de trabajo en las que seguir desarrollando, para así poder aportar nuevas ideas dentro del ámbito del seguimiento y análisis de medios de comunicación en Internet.

CAPÍTULO 2. ESTADO DEL ARTE

2.1 INTRODUCCIÓN

El nuevo milenio se presenta como la era donde el conocimiento y la información es el activo intangible más preciado por las organizaciones y se le ha dado a las Tecnologías de la Información y Comunicaciones (TIC), la tarea de gestionarlos y tratar de preservarlos de generación en generación a través de sistemas que permitan administrar, recopilar, organizar, analizar y diseminar este saber tratando de brindar la información adecuada a la persona correcta en el momento oportuno para proporcionar su creatividad, excelencia y competencia.

La recuperación de información es una actividad que el ser humano realiza, consciente e inconscientemente, casi continuamente, y en el marco de cualquier otra actividad. La necesidad de resolver una duda, o de documentar una afirmación o estudio, son expresiones clásicas de los procesos de recuperación de información. Con el desarrollo de los sistemas digitales de procesamiento de datos y de tratamiento de información, las técnicas de recuperación de información han ido desarrollando un conjunto de teoría y aplicación práctica que subyacen en la actualidad a cualquier actividad en entornos informáticos.

Con la excepcional expansión que ha sufrido Internet en los últimos años, se ha puesto a disposición de los usuarios un ingente volumen de información. El tamaño de la Web es gigantesco, y continúa creciendo según un modelo exponencial, siendo cada vez mayor el número de fuentes de contenidos y el volumen de datos disponible. Sin embargo, esto puede llegar a suponer un inconveniente, ya que tal sobrecarga de información, puede reducir notablemente su usabilidad. Por ello, uno de los grandes retos profesionales de la actualidad, dado la inmensidad de la red y su rápido crecimiento, es conocer el método idóneo para localizar la información que deseamos rápidamente [32].

El conocimiento de las características propias de los documentos (páginas web) que el usuario puede encontrar en Internet, y de la teoría de la recuperación de información, son los pilares básicos sobre los que construir una técnica adecuada de recuperación de información en Internet [26].

En este escenario, los buscadores de Internet han adquirido un enorme protagonismo y son la principal fuente de información para la mayor parte de los internautas. Hoy en día, las consultas en buscadores o directorios se presentan como la actividad más frecuente realizada en la Red, por encima de la descarga de software, el uso del correo electrónico, la participación en foros y chats o la descarga de contenidos multimedia [17].

Actualmente, un buscador de Internet consiste en una dirección más de la WWW que, mediante el uso de una base de datos, ofrece al usuario las direcciones URL de otras páginas o servicios, u otro tipo de informaciones en el caso de algunos buscadores especializados, atendiendo al criterio de búsqueda que se haya seleccionado. Pero además de ofrecer dichas direcciones, permiten, en la mayoría de los casos, acceder a los recursos localizados mediante enlaces, facilitando así la navegación al usuario, que no tiene más que conocer única y exclusivamente la dirección de su buscador favorito para moverse por el inmenso océano de Internet sin perderse.

En general los buscadores especializados son herramientas que restringen la búsqueda en la web a aquellos recursos que cumplen una serie de requisitos: tipo de documento (libros, artículos, etc.), materia (ciencia, humanidades, etc.) o nivel de la información (documentación de carácter científico y académico). El buscador desarrollado en el proyecto, como ya hemos comentado, se enfoca en la información de los principales medios de comunicación online, a partir de la descarga periódica de sus feeds RSS.

Si bien los periódicos han llegado quizás un poco tarde al fenómeno Internet y la mayoría lo siguen utilizando sólo como una biblioteca, el nuevo modelo de red, web 2.0, está pensado como una plataforma para crear, compartir y distribuir información proporcionando herramientas que permiten la integración de un tejido social, es decir, una red de personas que pueden interactuar a través de los espacios que se han generado en Internet, tales como blogs, google groups, twitter, facebook, wikipedia y un sinfín de útiles aplicaciones que permiten la interrelación de información [33].

De modo que la clave está en la manera de concebir el modelo de trabajo. Mientras que la web 1.0 funciona dando por sentado que el editor es el que se encuentra en el lado del servidor, y el cliente es el consumidor de información. Así, CNN [34] pone sus datos en CNN.com y yo voy a esa página y hago clic sobre las noticias que llamen mi atención. La web 2.0, cambia esta perspectiva, convirtiendo al público en el editor de la información.

Como ejemplos se pueden destacar los blogs, que se han convertido en medios de masas de alcance aun mayor para obtener información sobre intereses de mercado o noticias sobre lugares peligrosos. Los podcast, que han reducido los costes y las barreras tecnológicas. O la creciente popularidad del wiki, reflejada en el imparable crecimiento de Wikipedia [35].

Además de publicar sus propios contenidos, los usuarios se están volviendo “proactivos”, es decir, definen cómo desean ver el contenido de otras fuentes. Los grandes consumidores de información pueden encontrarla en RSS, recogidos y agrupados por un agregador, lo que deja entrever que en esta nueva fase de la red, el contenido es más importante que el continente.

Y bajo estos pilares se ha centrado el proyecto desarrollado, creando nuestro propio “mashup”, es decir, un sitio web que utiliza el contenido procedente de otras aplicaciones web, agrupado por categorías para su posterior análisis.

2.2 RECUPERACIÓN DE INFORMACIÓN

La dificultad de establecer una definición única sobre lo que es la "información", o sobre lo que el termino representa, resulta una paradoja si se considera que, precisamente ahora, es común afirmar que se vive en la era de la información y que es en estos momentos cuando aumentan en progresión geométrica las herramientas que el hombre tiene a su disposición para reconocer y manipular esa información [36].

En general, la información puede comprenderse como un conjunto de conocimientos o hechos derivados de datos, que son por naturaleza repetitivos y redundantes, que describen un mundo que consiste en procesos y eventos que ocurren una y otra vez con pequeños cambios. La información puede ser tangible o intangible, pero siempre reducirá la incertidumbre sobre un estado o suceso.

Además, es importante hacer una distinción entre datos, información y conocimiento (**Figura 1**), sin entrar en mucho detalle, nos bastará con comprender que la información, en sí, supone algún tipo de transformación sobre la masa de datos sin analizar. Se suele aceptar que la información es el resultado del proceso de datos. Y que se produce conocimiento cuando el usuario asimila la información, la contextualiza en su entorno y adquiere un saber que le es de utilidad en una situación dada.

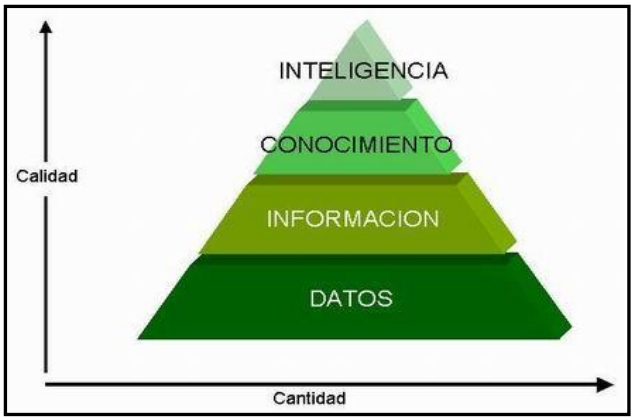


Figura 1 — Pirámide de la información

El proceso de recuperación de información (RI) tiene una motivación, unas fases, unas técnicas e instrumentos y unos resultados. Se inicia un proceso de recuperación de información cuando una persona detecta una inconsistencia o carencia en su estado de conocimientos que le impide tomar una decisión o desarrollar una acción. Los investigadores de la recuperación de información, por ejemplo Belkin [3], han enfatizado en este punto, señalando la presencia de un estado anómalo de conocimiento (ASK, *Anomalous State of Knowledge*), o de un problema de incertidumbre, de ajuste del espacio mental del individuo (*problem space*), como Ingwersen [15].

Cuando el individuo siente esa inquietud, comienza a desarrollar un conjunto de acciones que persiguen devolver el equilibrio a sus esquemas mentales. Estas acciones se centran en la búsqueda de nueva información, y suelen estar mediadas por un proceso de comunicación, que puede ser una interacción persona-persona, o bien una interacción hombre-máquina (HCI, *Human Computer Interaction*).

En realidad, los procesos de recuperación de información suelen ser bastantes continuos, y generalmente la resolución de un problema trae como consecuencia la aparición de otros nuevos. Se puede concluir por tanto, que la RI es el conjunto de procesos envueltos en la representación, almacenamiento, búsqueda y localización de información que es relevante para resolver un requerimiento formulado por un ser humano.

El complejo proceso de RI engloba numerosas tareas, de las que la consulta de recursos de información electrónica resulta ser una más de ellas [27]. El auge que están teniendo en los últimos años los sistemas de información de todo tipo, desde las Administraciones Públicas hasta las pequeñas y medianas empresas con sus sistemas de información contable, han favorecido que la mayor parte de las actividades relacionadas con la búsqueda y localización de la información se desarrollen sobre sistemas informáticos.

2.3 MODELOS DE RECUPERACIÓN

La recuperación de información necesita de unos algoritmos y modelos especializados para conseguir sus fines. Estos modelos tienen como objetivo el facilitar el proceso de comparación entre una consulta determinada y un conjunto de textos sobre los que se realiza la consulta. La principal clasificación para los modelos de recuperación de información es la siguiente (**Figura 2**) [38].

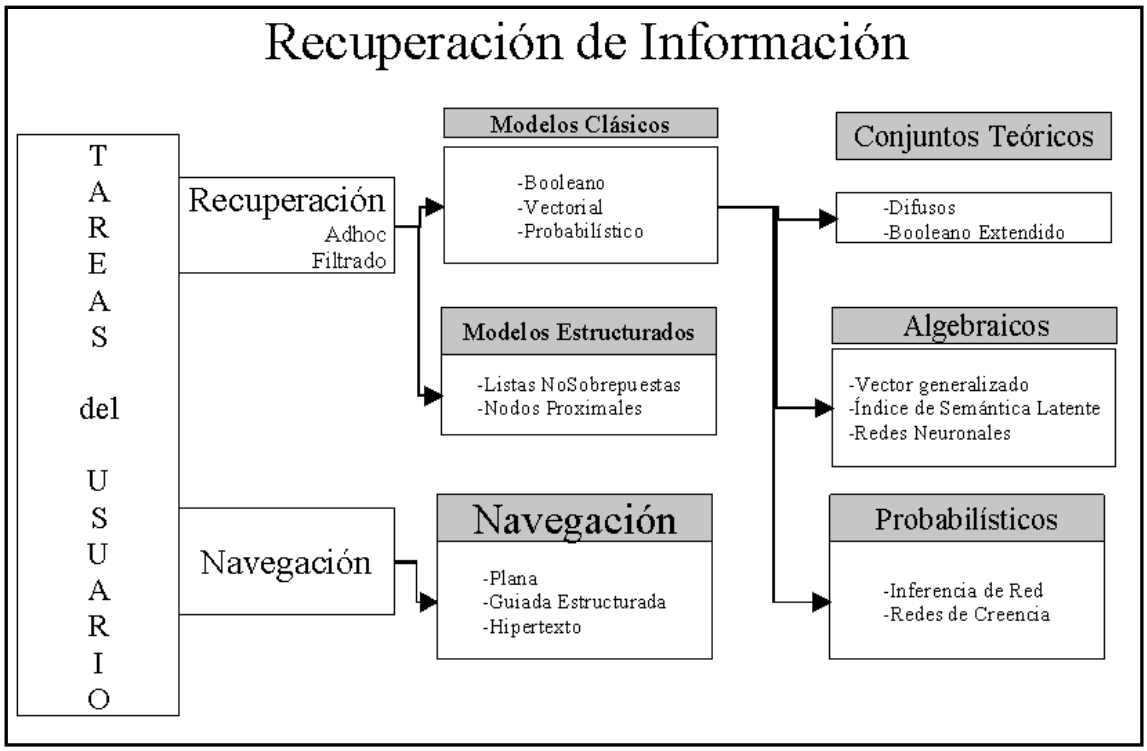


Figura 2 — Clasificación de los modelos de recuperación

Modelo Booleano

Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Dada su inherente simplicidad y su pulcro formalismo ha recibido gran atención y ha sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (relevante o no relevante).

Cuando se efectúa una consulta de una palabra sobre un documento, el sistema etiqueta los documentos como relevantes con la condición que contengan dicha palabra.

Desafortunadamente, el modelo booleano arrastra los mayores inconvenientes [2] de todos los modelos, entre algunos podemos destacar: no diferenciar entre documentos más o menos relevantes, sin ninguna posibilidad de mantener una escala gradual; no tener en cuenta el número de repeticiones de una palabra en el documento; o no permitir ordenar los resultados por ningún criterio.

Para facilitar la evaluación de las consultas sobre la base documental, se utilizan habitualmente representaciones en forma de árbol de los documentos relevantes como se puede observar en la **Figura 3**. De este modo se recorre el árbol de abajo a arriba determinando el conjunto de documentos relevantes.

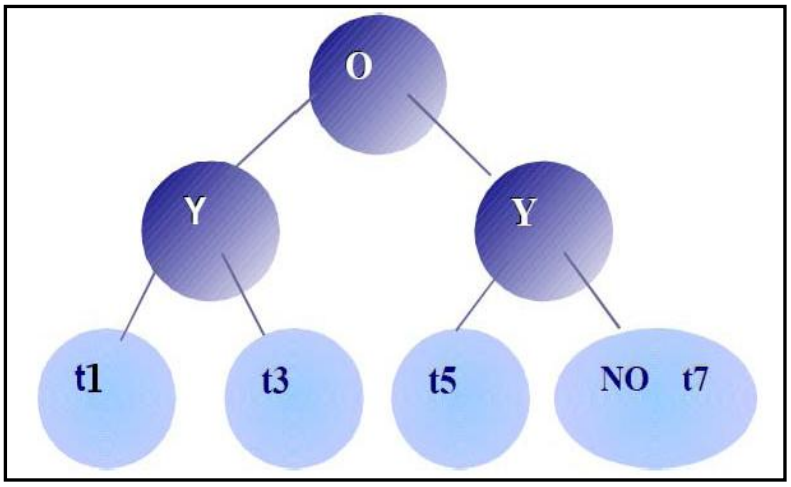


Figura 3 — Representación en forma de árbol de la consulta: (t1 y t3) o (t5 y NO(t7))

Modelo Vectorial

El modelo vectorial reconoce que el uso de medidas binarias es altamente limitador, y propone un marco de trabajo en el cual sea posible la coincidencia parcial [2]. Partiendo de que se pueden representar los documentos como vectores de términos, los documentos podrán situarse en un espacio vectorial de n dimensiones, es decir, tantas como elementos

tenga el vector. Cada documento está entonces en un lugar determinado por sus coordenadas, al igual que en un espacio de tres dimensiones cada objeto queda bien ubicado si se especifican sus tres coordenadas espaciales. Se crean así grupos de documentos próximos entre sí debido a sus similitudes.

Es el modelo de recuperación, más popular y utilizado hoy en día, esto se debe a sus principales ventajas como son: el esquema que utiliza; su estrategia de coincidencia parcial; su fórmula de clasificación; o su simpleza entre otros. Sin embargo, no incorpora la noción de correlación entre términos.

Modelo Probabilístico

El modelo probabilístico se basa en la mejora del rendimiento utilizando la información procedente de la distribución estadística de los términos, de manera que la frecuencia de aparición de un término en un documento o conjunto de documentos podría considerarse un dato relevante a la hora de establecer una consulta a la base de datos documental.

La principal ventaja de este modelo es que, en teoría, recupera de forma ordenada los documentos que con mayor probabilidad son relevantes. Sin embargo, entre las desventajas podemos citar: la necesidad de suponer la separación inicial de los documentos en dos conjuntos, los relevantes y los no relevantes; o el hecho de que este método no tiene en cuenta la frecuencia con la cual un término aparece dentro de un documento [23].

Modelos Estructurales

Combinan la información del contenido del texto con la información sobre la estructura del documento [39]. Existen dos tipos: el método de las listas no sobrepuestas, que divide el texto en regiones *no superpuestas* las cuales son coleccionadas en una lista, para implementarlo se crea un archivo invertido en el que cada componente estructural es una

entrada en el índice y asociado con cada una de estas entradas, hay una lista de regiones de texto como una lista de ocurrencias **(Figura 4.1)** [38].

Y el método de los nodos proximales, que define estructuras de indexación jerárquica independientes sobre el mismo texto, para su implementación primero se buscan los componentes que coinciden con la cadena especificada en la consulta, y a continuación se evalúa cuál de estos componentes satisface la parte estructural de la consulta **(Figura 4.2)** [38].

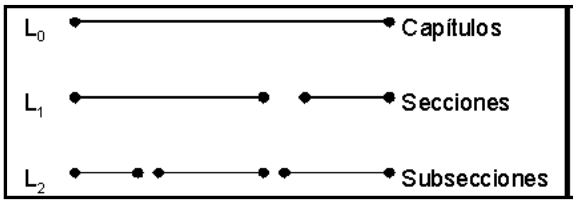


Figura 4.1 — Modelos listas no sobrepuestas

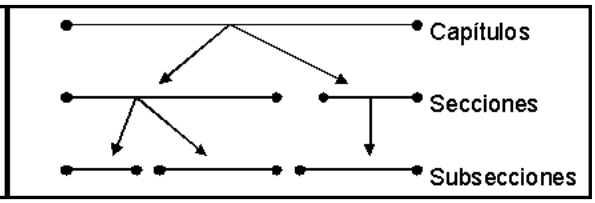


Figura 4.2 — Modelos nodos proximales

2.4 LA RECUPERACIÓN DE INFORMACIÓN EN INTERNET

La teoría de la RI en Internet parte de los logros y limitaciones de la teoría de recuperación de información, desarrollados desde la década de 1960. Tan temprano como comenzó el crecimiento geométrico de la información en Internet comenzaron a desarrollarse instrumentos para facilitar la localización y acceso a los recursos de información [13]. Cabe recordar aplicaciones como Archie (para localizar ficheros en servidores FTP), Whois, NetFind o Veronica entre otros.

El desarrollo acelerado de Internet y la necesidad de desarrollar herramientas que facilitasen la localización y acceso a la información en Internet obligó a adoptar, en un primer momento, dos aproximaciones clásicas similares a las ya existentes en los entornos de documentación automatizada, y que a día de hoy son la base de los buscadores en Internet:

- La creación de listados, índices y catálogos ordenados por áreas o materias, de forma que el usuario dispusiese de un conjunto de fuentes seleccionadas en

que empezar a buscar. El ejemplo más conocido es el de Yahoo! [40], aunque en la actualidad también combine un motor de búsqueda.

- La creación automática de bases de datos basadas en índices o ficheros inversos mediante unas aplicaciones que rastrean o exploran todo el ámbito web, llamados robots o spiders. El ejemplo más conocido es el de Google [41].

Pero además de las innumerables ventajas que supone el proceso de recuperación de información en Internet, también tenemos que abordar las limitaciones que existen derivadas de la estructura hipertextual o de la dinámica de actualización de documentos, o bien a factores externos a las técnicas de recuperación de información y al propio usuario:

- Existe una disfunción entre los procesos de indización automática, la representación del contenido de un documento que se crea y el contenido real.
- La cobertura de los motores no es exhaustiva y además el contenido de los diferentes motores se solapa en parte.
- La actualización de las bases de datos no es automática. Las variaciones que puede sufrir una página web no son automáticamente reflejadas en los motores.
- Las respuestas que ofrecen las herramientas a las ecuaciones formuladas no presuponen fiabilidad ni rigor. Sólo responden a la ecuación planteada. La consideración de si los resultados obtenidos merecen fiabilidad y confianza queda a discreción del usuario.

Una de las claves para una utilización efectiva de las herramientas que provee la recuperación de información, es conocer el tipo de información que se está publicando en la web [30]. Cualquier tema es tratado en Internet, sin embargo también es cierto que existe mucha información de mala calidad “ruido”, que dificulta la búsqueda.

Por todo ello es fundamental comenzar acotando la búsqueda a realizar tan estrechamente como sea posible, identificando los parámetros siguientes: el tema de búsqueda; los límites geográficos, con objeto de elegir el tipo de buscador; el idioma; el objetivo de la búsqueda; el tipo de fuente donde puede encontrarse (pagina web, grupos de noticias, bases de datos, etc.), entre otros.

Sin embargo, a pesar de que los usuarios de la red no tengan claro estas pautas y de la presencia entre los resultados de "ruido", el uso de las herramientas de búsqueda, es hoy en día la actividad más frecuente realizada en la red, siendo Google el buscador más utilizado por los internautas [42] y [43].

2.5 TIPOS DE BUSCADORES

No todos los buscadores son iguales en su funcionamiento ni en sus posibilidades. Algunas de sus características son estructurales y, por tanto, básicas, mientras que otras los hacen diferir menos significativamente [1].

1. Según su base de datos

Los buscadores generan las bases de datos en las cuales se efectúan las búsquedas de dos maneras distintas. Éstas dan lugar a los dos tipos de buscadores existentes: los índices y los motores de búsqueda. La primera consiste en organizar los datos en categorías, de modo que toda la información quede clasificada según los criterios temáticos definidos por el administrador del buscador. La segunda forma de generar la base de datos es mediante el uso de ciertos programas exploradores denominados robots, y más concretamente arañas.

2. Según la información que localizan

Existen buscadores en los que todas las direcciones posibles, cualquiera que sea el tema al que pertenezcan, tienen cabida. En el caso de los índices, porque puedan situarse en

alguna de las subcategorías, y en el de los motores, porque no se distinga entre las mismas. Estos son los buscadores más amplios, pues su información es general, pero también encontramos otros cuyo interés se centra en un solo tema, como puede ser el software, la física o las imágenes, por citar algunos ejemplos. Estos buscadores están experimentando un gran auge, dado que existe en la red un interés mayor, por parte de ciertos grupos de usuarios, en algunos temas concretos, como la ciencia, la programación, el software, la música, etc.

Y también hay servicios que no se encargan de localizar direcciones URL, como en los dos casos anteriores, sino direcciones de correo electrónico de usuarios de la red, entre otras informaciones todas ellas personales.

3. Según el número de bases de datos a las que acceden

Los primeros buscadores disponían cada uno de ellos de una única base de datos, pues eran servicios independientes unos de otros. Pero en la actualidad han aparecido servicios que efectúan la consulta del usuario en diversas bases de datos distintas, ofreciendo así unos resultados de posibilidades más amplias.

- **Una única base de datos**: a este tipo pertenecen los buscadores de primera generación. La información se localiza a partir de su propia base de datos, por lo que siempre sufrirá de las deficiencias que ésta presente.
- **Varias bases de datos no simultáneas (multibuscadores)**: para mejorar las prestaciones de buscadores de una única base de datos, nacen los de tipo múltiple. La información en ellos se trata de localizar en diversas bases de datos diferentes, aumentándose las posibilidades de encontrarla. Los multibuscadores realizan esta operación manualmente, pues ponen varias bases de datos a disposición del usuario, pero es él mismo quien debe elegir cuál de ellas desea utilizar en cada momento.

- **Varias bases de datos simultáneas (metabuscadores)**: son buscadores múltiples que llevan a cabo la consulta solicitada por el usuario en varias bases de datos al mismo tiempo.

4. Según el ámbito geográfico

La región geográfica de un buscador condiciona la información disponible en su base de datos cuando su interés principal es la inclusión de páginas pertenecientes a dicha zona concreta. De este modo, podemos encontrar buscadores que disponen de direcciones de todo el mundo, de continentes, de países, de regiones, etc.

2.5.1 Directorios

Para mejorar la eficacia en la exploración de este gran espacio de información que constituye la World Wide Web, una posibilidad es dividirlo en diferentes categorías temáticas significativas para los usuarios [5].

Las bases de datos de estos buscadores se diseñan para permitir que la información pueda agruparse en distintos temas, siguiendo un orden jerárquico, dentro del cual las direcciones se disponen de más generales a más concretas. Esto es equivalente a crear un árbol cuyas ramas principales son los temas, y que se ramifican en subtemas, o grupos de menor categoría. En este caso es el administrador del buscador quien, mediante la inscripción en el mismo de las direcciones URL, forma la base de datos, que se va actualizando de forma permanente por un equipo que visita las direcciones y por las altas de los usuarios.

Entre las ventajas de los directorios se encuentra, además de la calidad de las web indexadas, la posibilidad de hacerse una idea de los sitios más relevantes sobre un determinado tema. Dos de sus principales inconvenientes son su lentitud y su reducido catálogo, si los comparamos con un motor de búsqueda. Además, en la mayor parte de los

directorios hay un alto índice de páginas que ya no existen y que continúan siendo listadas porque no han sido dadas de baja.

A la hora de buscar en un directorio hay que tener también en cuenta que, en numerosas ocasiones, el orden en que están colocados los enlaces responde a un criterio comercial. Hay empresas que pagan porque su web ocupe los primeros lugares, ya que está demostrado que de esta forma aumenta su tráfico.

Por otro lado, los servicios de consulta basados en directorios han incorporado cada vez más prestaciones convirtiéndose en una puerta de acceso a todas las posibilidades que ofrece Internet. Esta evolución ha dado lugar a lo que, hoy día, se denominan portales. Un portal es un conjunto de servicios que pretenden satisfacer todas las necesidades del navegante, aunque sea bastante difícil ajustarse a la demanda de los millones de usuarios potenciales. Los mejores directorios generales de la red juegan sus grandes bazas en dos de las mayores áreas de demanda de los internautas actuales: las noticias en línea y cuestiones de finanzas personales, pero además también ofrecen acceso a compra en línea y a variados servicios como la bolsa, el tiempo, correo electrónico, chat, etc.

Ejemplos de este tipo de buscadores son Yahoo! [40], Open Directory [44] o Galaxy [45]. Sin embargo, los buscadores basados en directorios han evolucionado y ahora ofrecen, casi siempre, un motor de búsqueda en su página principal.

2.5.1.1 Yahoo! (http://www.yahoo.com)

Yahoo! [40] se creó en abril de 1994 por iniciativa de dos estudiantes de la universidad de Standford, David Filo y Jerry Yang. Yahoo significa *"Yet Another Hierarchical Officious Oracle"*. Los empleados de Yahoo! se encargan de examinar páginas web y recursos de todo el mundo y los incluyen en esta guía temática universal. Una vez examinado el material,

incorporan cada una de las páginas en una categoría predeterminada, hacen un pequeño resumen de su contenido y lo publican en el catálogo general.

Aunque Yahoo! siempre ha sido considerado como un directorio, en diciembre de 2002 compró el potente motor de búsqueda japonés Inktomi. El objetivo era apartar de sus páginas a Google, que hasta entonces era su motor de búsqueda, y entrar en la lucha por posicionarse entre los mejores puestos en el mercado de la búsqueda de información por Internet. Además, en 2003 adquirió el servicio de resultados de pago Overture, que tenía en ese momento acuerdos para servir sus listados a MSN Search y AOL Europa, entre otros. En 2009, para derrocar al gigante de las búsquedas (Google), Microsoft y Yahoo anunciaron un acuerdo en el que Bing [46], motor de búsqueda de Microsoft, sería propiedad de Yahoo! Search, mientras que Microsoft podría utilizar durante 10 años las tecnologías de búsqueda de Yahoo!.

2.5.2 Motores de búsqueda

Los motores de búsqueda son el tipo de buscador web más utilizado y extendido en la actualidad. Todos presentan una estructura similar constituida principalmente por la base de datos, el programa de indización, el robot de búsqueda (por la automatización con que realiza su labor de rastreo) o araña (Web significa tela de araña en inglés, así que algo que se mueve por la Web es fácil que sea una araña) y la interfaz. Sin embargo, los antecedentes de los buscadores de hoy fueron simplemente listados de direcciones de recursos y documentos electrónicos de la red que alguien, por iniciativa particular o institucional, pensaba que podían ser de interés para otros internautas. Un repaso de la historia de cada buscador permite comprobar cómo, en su mayor parte, comenzaron como proyectos de investigación o incluso como divertimento, de estudiantes graduados, profesores de universidad, ingenieros, programadores de sistemas, etc.

2.5.2.1 Módulos de un buscador

Conocer el funcionamiento interno de los motores de búsqueda es importante para llegar a comprender el mundo que hay dentro de la Red (**Figura 5**). Además, cuando un usuario utiliza un buscador, debe ser consciente que no está buscando en toda la World Wide Web, sino en la base de datos específica del motor [24].

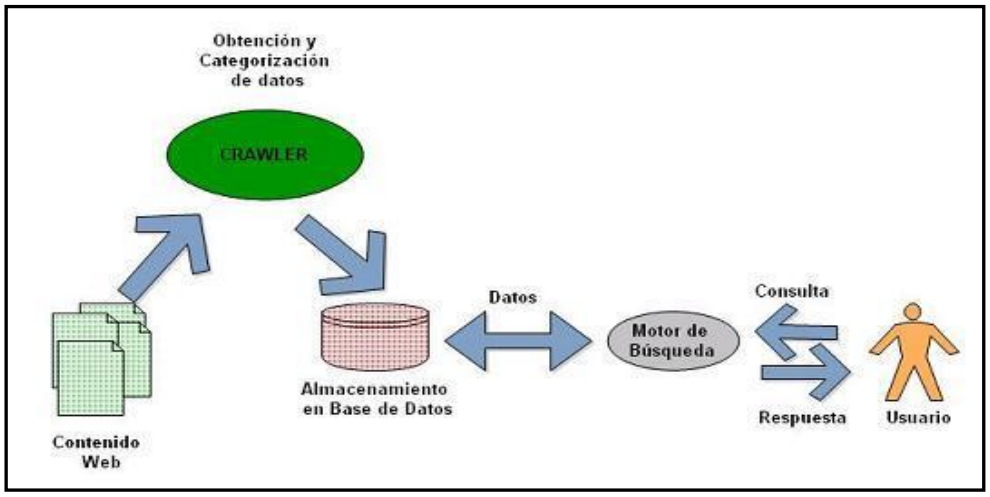


Figura 5 — Estructura de un motor de búsqueda

La araña

Es un pequeño programa que parte con una lista inicial de URL o direcciones web. Una vez en esas páginas, va siguiendo los enlaces que vinculan unas páginas con otras y cada vez que encuentra una página web que el sistema no tiene en su listado, hace una copia del contenido de la página y la almacena en la base de datos. El robot actualiza la base de datos visitando periódicamente las páginas para comprobar si ha habido alguna modificación o si aún siguen activas, incluso, pueden "aprender" a visitar y reexaminar con mayor frecuencia aquellos servidores que cambian rápidamente o que son citados en otras muchas páginas.

Cada motor de búsqueda comercial tiene su propia implementación de este componente, así la araña de Google se llama Googlebot [56] o la de Yahoo! se identifica como Slurp; y es en este aspecto fundamental donde reside gran parte del éxito de cada motor.

En general, la calidad de estas arañas depende de los siguientes aspectos [21]:

- La eficacia para descubrir nuevos documentos y mantener la lista de direcciones.
- La cantidad de información que guarda el fichero invertido sobre cada documento.
- La potencia del lenguaje de consulta.

Si bien, a pesar de las ventajas que proporcionan, los robots por su condición y finalidad, pueden alterar en gran medida el equilibrio en la red, consumiendo excesivos recursos y creando demasiadas dificultades a los administradores de los servidores. Poniéndose así de manifiesto la necesidad de regular ciertos aspectos en la www, como la moderación del consumo de recursos y la compartición de resultados, o como informar a los servidores de enlaces obsoletos.

La base de datos

Está constituida principalmente por un índice de palabras, frases y datos asociados a la dirección de cada recurso (URL), aunque en la actualidad también incorporan programas, imágenes, archivos, etc. La lista de elementos indizados en la base de datos varía de una herramienta de búsqueda a otra. Algunas indizan cada palabra de las páginas web, incluyendo la URL y el texto de algunas metaetiquetas como *author, title, keywords o description*. Esta información puede mejorar sensiblemente la eficacia en la recuperación y en la ordenación de los resultados o ranking. Otros indizan únicamente las palabras de aparición más frecuente, o las incluidas en ciertas etiquetas, o sólo las primeras palabras o líneas de los documentos HTML.

El indexador

Es el sistema de gestión de bases de datos. Las herramientas de búsqueda disponibles en Internet utilizan distintos métodos para indizar los recursos que incorporan a sus bases de datos. Por ejemplo, para facilitar la búsqueda, algunos motores de búsqueda eliminan las palabras denominadas *"stop words"* o palabras vacías, como pueden ser: preposiciones, conjunciones, artículos, etc.

Hay que tener en cuenta que las bases de datos tienen almacenados los contenidos de una determinada web en el momento concreto en que esa página fue indexada, por lo que puede ser que ésta esté desactualizada cuando, después de un tiempo, sea el resultado de una búsqueda.

Interfaz de recuperación

Ésta es la parte más compleja del motor. Ésta compuesta por varios módulos que incluyen: el formulario de búsqueda, la máquina que evalúa una búsqueda y la hace coincidir con los documentos más relevantes en la base de datos en la que están indexadas las páginas, y los resultados de dicha búsqueda.

El formulario y los resultados varían poco de un motor a otro. Todos tienen modelos de búsqueda simples y avanzados, y la mayor parte de los resultados son similares (**Figura 6**), como se observa en la gráfica comparativa entre los resultados que ofrecen Google y Yahoo! para una misma palabra "uc3m", con algunos añadidos como pueden ser webs relacionadas con las paginas encontradas, las búsquedas más populares, etc. [47].

Lo que realmente es más significativo de un motor es la forma en que éste calcula la relevancia, es decir, la importancia que se da a una determinada web y que influirá en el orden en el que aparecerá en la lista de resultados del usuario. Algunos de ellos se basan en análisis estadísticos del texto y desarrollan sofisticados métodos de comparaciones en su tarea de

localizar los documentos más relevantes a una búsqueda. Otros calculan la relevancia según el número de enlaces que hay en la Red apuntando a la página web que se trata de analizar.

Estas fórmulas son el secreto mejor guardado de las compañías dueñas de las maquinas de búsqueda, y sus algoritmos son actualizados constantemente para mejorar la calidad o incorporar las últimas tecnologías contra los "*spammers*", es decir, aquellos que se dedican a incluir basura en los buscadores y que entorpecen el buen funcionamiento.

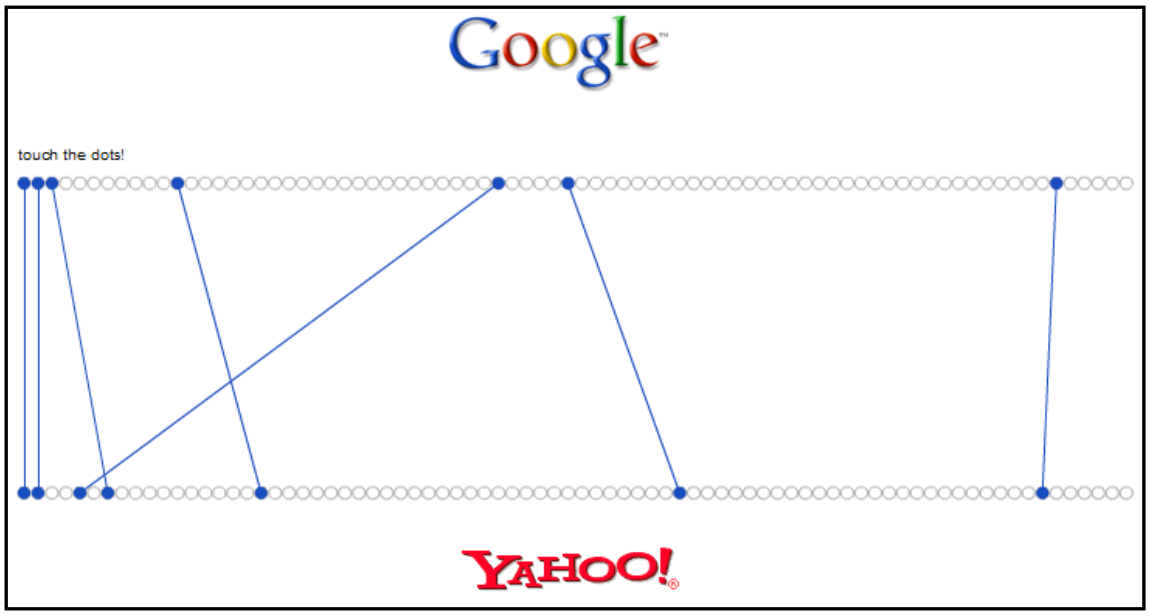


Figura 6 — Comparativa entre "Google" y "Yahoo": Keyword "uc3m"

2.5.2.2 Google (http://www.google.com)

Google [41] nació el 15 de septiembre de 1998. Sergey Brin y Larry Page, de la Universidad de Standford (Estados Unidos) estaban trabajando en un proyecto de clase para identificar patrones en la estructura de enlaces de la Red. Fue este estudio lo que les dio pie a diseñar un motor de búsqueda basado en la estructura de los enlaces [24].

En la actualidad, Google es unos de los mejores motores de búsqueda que hay, siendo capaz de reconocer e indexar documentos en formato PDF, RTF, PostScript, Word, Excel y PowerPoint, entre otros. Cada 28 días, indexa 3000 millones de documentos web, y actualiza más de tres millones de páginas web importantes cada día.

Google marcó el nacimiento de una nueva generación de buscadores capaces de rastrear de forma automática todo el contenido de cada página Web y tener en cuenta factores ajenos a la propia página, de alguna forma más objetivos, para generar un sistema de consulta capaz de ofrecer resultados ordenados por su distinta relevancia en un grado jamás logrado hasta la fecha.

Tecnología de Google

La exitosa tecnología del buscador Google puede ser resumida en estos tres puntos [24]:

- **Algoritmo PageRank**: este algoritmo es uno de los factores que Google tiene en cuenta para establecer la importancia de una página web determinada. La filosofía sobre la que se sustenta este concepto es atribuir mayor importancia o credibilidad a una página web cuanto mayor número de otras páginas web contengan un enlace que apunta hacia ella.

  El concepto de PageRank de Google fue revolucionario, en el sentido de que por vez primera era capaz de incluir conceptos de valoración ajenos al contenido de la página.
- **Robot de búsqueda de Google**: Googlebot es el nombre que recibe el robot de búsqueda que rastrea Internet con el objetivo de indexar y almacenar el mayor número posible de enlaces web.
- **Uso de Linux**: la mayor parte de los servidores de Google utilizan Linux como sistema operativo, concretamente la distribución RedHat. Además, para el desarrollo de Googlebot y el resto de aplicaciones, los programadores de Google utilizan lenguajes como C o C++ y herramientas y entornos de desarrollo como XEmacs, Gcc, Gdb, Gnats, entre otras. Tal y como confiesan los creadores de Google, este sistema operativo fue seleccionado frente a

otros por su inmejorable relación rendimiento-precio y por el grado de personalización que ofrece.

También cabe destacar que, independientemente de la búsqueda de enlaces web, Google mantiene accesible para todos sus usuarios su laboratorio, en el que se encuentran disponibles muchos de los proyectos en los que esta compañía está trabajando, entre ellos podemos destacar: Google Video, Google Suggest, Google Scholar, o Google Desktop Search.

2.5.2.3 MSN Search (http://www.bing.com)

Microsoft no se ha querido quedar fuera de esta batalla por convertirse en líder en la búsqueda de información por Internet y a principios de 2005 lanzó al mercado su propio motor de búsqueda, reemplazando así la tecnología de Yahoo! que utilizaba hasta entonces en su sitio de MSN. En 2009, el motor de búsqueda de Microsoft pasó a llamarse Bing, el cual según la compañía, selecciona sus resultados a partir de una base de datos de más de 5000 millones de documentos y páginas web [24].

Una de las grandes ventajas del MSN Search es que ofrece un servicio de respuestas directas procedentes de los más de cuarenta mil artículos de su enciclopedia virtual Encarta.

2.5.3 Metabuscadores

Los metabuscadores permiten realizar una búsqueda en varios buscadores a la vez. Uno de sus inconvenientes, además de un mayor tiempo de espera, es que no suele ser posible precisar la búsqueda, ya que cada uno de los motores que engloba tiene sus propias peculiaridades de búsqueda. Su método de funcionamiento es el siguiente: cuando el usuario realiza una búsqueda, el metabuscador la dirige a sus motores asociados, componiendo una lista de aciertos que representan, teóricamente, las mejores respuestas a la pregunta.

Posteriormente, algunos ofrecen la posibilidad de ordenar por relevancia, entendiendo por relevancia el grado con que la web resultante concuerda con la búsqueda realizada por el usuario o con la información que necesita.

Algunos ejemplos de metabuscadores más famosos que podemos encontrar en la red son: Ixquick[1], MetaCrawler[2], Search Caddy[3], o All The Web[4].

2.5.4 Anillos Web

Los anillos web (*webrings*) son grupos de sitios web relacionados con un mismo tema. Cada web tiene insertado en sus páginas un código de programación gracias al cual aparecen agrupadas distintas webs por temas. Existen anillos sobre informática, mercadotecnia, medicina, videojuegos, alpinismo, etc.

Algunos sitios web que ofrecen este método de búsqueda son los siguientes: WebRing[5], The Rail[6], o Looplink[7] entre otros.

2.5.5 Portales horizontales y verticales y canales temáticos

Los portales son sitios web que actúan como distribuidores a otros sitios vinculados a un tema o campo. Se distinguen dos tipos: horizontales y verticales. Los primeros ofrecen información y servicios a un universo amplio. Mientras, los portales verticales, o “vortales”, ofrecen información específica y seleccionada sobre un tema.

Por otro lado los canales temáticos se centran en materias específicas que suelen interesar a un amplio número de usuarios. Se diferencian de los portales verticales principalmente en que estos últimos tienen una dirección común a varios de ellos y suelen integrarse en un portal generalista, mientras que los canales temáticos son independientes empresarialmente hablando.

1: www.ixquick.com; **2:** www.metacrawler.com; **3:** www.searchcaddy.com; **4:** www.alltheweb.com
5: www.webring.com; **6:** www.therail.com; **7:** www.looplink.com

Se diferencian de los portales verticales principalmente en que estos últimos tienen una dirección común a varios de ellos y suelen integrarse en un portal generalista, mientras que los canales temáticos son independientes empresarialmente hablando.

Ante la cantidad tanto de portales horizontales y verticales como de canales temáticos, merece la pena citar como recursos algunos de los buscadores y directorios de portales más utilizados como son: PortalCual[1], Portalmix[2], o Buscadores y Portales[3].

2.6 LOS BUSCADORES DEL FUTURO

En un futuro cada vez más próximo, la Web estará dotada de cierta "inteligencia artificial", en cuyo desarrollo se lleva trabajando años, y que responde al nombre de Web semántica. Se trata de lograr que la información disponible en Internet no se componga de meros datos sin sentido alguno para los ordenadores, sino que éstos logren, en cierta medida, "entenderla", con el objetivo de automatizar cuantas más tareas posibles [24].

Una de las opciones que ofrecerá al usuario la Web semántica será la posibilidad de realizar preguntas complejas en un buscador; éste responderá directamente a la pregunta realizada, sin recurrir a un listado de webs aparentemente relacionadas como ocurre en la actualidad. Por ejemplo a la pregunta de quién descubrió América, el buscador devolverá mediante un visualizador: "Cristóbal Colón", y acompañará la respuesta con los enlaces de referencia que lo confirmen.

Según el investigador de la Universidad Autónoma de Madrid Pablo Castells, desarrollador de diversos proyectos basados en las tecnologías de Web semántica, dotar a la Web de "inteligencia" será muy útil para cambiar la forma en la que el usuario se comunica con los distintos servicios disponibles en Internet. "Podrá expresar consultas de una forma natural y a la vez obtener respuestas más concretas".

1: www.portalcual.com; **2:** www.portalmix.com; **3:** www.buscadoresyportales.com

Uno de los proyectos desarrollados, es el proyecto Neptuno, que ha conseguido integrar esta tecnología en la hemeroteca digital del periódico de Lleida Diari Segre, con la consiguiente mejora significativa de las búsquedas de datos complejas.

Por otro lado, en el panorama internacional sobresale el proyecto mSpace, desarrollado por un grupo de investigadores de la Universidad de Southampton (Estados Unidos), que definen como un modelo interactivo que ayuda a establecer relaciones en la información. El proyecto consiste en un buscador semántico de música clásica, que combina la facilidad de manejo de la tienda de música en línea de iTunes con la información que proporcionaría una búsqueda en Google.

Mientras, la Universidad de Standford (Estados Unidos), ha desarrollado un buscador, con el nombre de Search on TAP, que utiliza las tecnologías y técnicas de Web semántica. Gracias a ellas añade información a los resultados por palabras clave, que despliegan los buscadores convencionales como Google o Yahoo!, que permite no sólo una mayor precisión en la búsqueda, sino una búsqueda “inteligente” [24].

De la misma manera los grandes motores de búsqueda que existen en la actualidad, conscientes de las ventajas que ofrecen las tecnologías de Web semántica, intentan hacer evolucionar su forma de localizar la información hacia una búsqueda semántica.

Por ejemplo, Google tiene ya un servicio que ofrece respuestas directas a algunas preguntas en la cabecera de la página de resultados. Su intención es evitar a los usuarios tener que navegar por las webs que pueden contener o no lo que el usuario busca, así si un internauta busca “Población España” o “Clima Madrid”, Google.es devolverá la respuesta directa junto a un enlace a la página de donde proviene esta información [48].

Otra mejora a los resultados de búsqueda que ofrece Google, se conoce como "fragmentos de texto enriquecido". Cuando sólo una sección de la página es relevante para la búsqueda, se incluyen enlaces que nos llevan directamente a las secciones específicas de la página donde está la información que buscamos siendo más fácil y rápido encontrar lo que queremos [49].

2.7 MEDIOS DE COMUNICACIÓN EN INTERNET

Hoy en día, los medios de comunicación constituyen una herramienta persuasiva que nos permite mantenernos en continua comunicación con los distintos sucesos sociales, políticos y económicos tanto a escala nacional como internacional.

Es tal la rapidez y fuerza con que los MCM (Medios de Comunicación de Masas) han ido incorporándose en nuestra realidad, que no ha dado tiempo ni para adaptarse. Si a esto se le añade que Internet ha generado grandes cambios en la comunicación social, ofreciendo la posibilidad de compartir, crear, generar o difundir contenidos, informaciones y datos desde cualquier lugar en cualquier momento, es normal, que los medios tradicionales abran sus puertas a la red y a los servicios que la web 2.0 puede proporcionar.

Este poder que la tecnología otorga, ha influido en la forma de generar y consumir información, permitiendo a los internautas tener su propio espacio de expresión. Entre las herramientas que permiten crear estos nuevos servicios, destacan los blogs, los wikis y los portales de networking social (red social).

2.7.1 Los blogs

Los blogs o weblogs [33], son pequeños mundos de información temática, abiertos al libre pensamiento de sus creadores y a la curiosidad de sus posibles lectores. Se trata de una especie de "diario de bitácora" en línea que aprovechan las grandes posibilidades técnicas de

la Red para aportar las más compleja visión sobre los más diversos temas, de forma que puedan constituir un recurso muy útil para alguien interesado en un tema específico.

Los blogs han entusiasmado a los usuarios por la rapidez de creación, la sencillez de uso y de actualización. En la actualidad, el blog es una función líder de la red. Se encuentran ya guías y publicaciones dedicadas a este fenómeno, agencias de comunicación especializadas en la promoción de los blogs y agencias publicitarias que garantizan la venta de espacios.

Sin embargo, en el principio, el blog era un diario personal, en el que, cuando lo deseaba el propietario ponía un post (nota) abierto a la consulta de los internautas, que, a su vez, introducían comentarios (sujetos a la moderación por parte del propietario del blog), añadían vínculos, etc. El orden de los posts era cronológico y haciendo clic sobre una fecha, aparecían las notas correspondientes a ella.

Ahora el blog se diferencia por su forma: fotoblog, videoblog y moblog (cuando el blog se actualiza a través del móvil) y su temática.

Algunas de las herramientas más utilizadas para la creación de blogs son Blogger[1], Movable Type[2], Antville[3] y Pitas[4]. En cualquiera de ellas, tan sólo es necesario elegir un nombre de usuario y una contraseña, escribir el artículo y publicarlo.

2.7.2 El wiki

Es un sitio de Internet [33] con contenidos que los visitantes pueden modificar, ampliar y añadir con total libertad. Los wikis se utilizan para facilitar la escritura colaborativa de documentos con las mínimas restricciones. La palabra wiki procede del término hawaiano “wiki wiki” que significa rápido o informal.

1: www.blogger.com; **2:** www.movabletype.org; **3:** www.antville.org; **4:** www.pitas.com

El más conocido y más visible de todos los wikis es la enciclopedia Wikipedia [35], que se ha convertido en una referencia para la búsqueda de información general. En parte gracias a Wikipedia, los wikis han alcanzado un nivel de respetabilidad y notoriedad que permite asociarlos a las Web 2.0.

A diferencia de los blogs, donde el propietario, ya sea una persona o una asociación, gestiona enteramente el contenido, en un wiki, todo el mundo tiene los mismos derechos de modificación sobre el contenido.

2.7.3 Las redes sociales

Son comunidades de personas que [24], usando la Red como medio, se relacionan entre sí. A partir de la teoría de Stanley Milgram, según la cual sólo hay seis niveles de distancia entre una persona y cualquier otra persona en cualquier otra parte del mundo, las redes sociales proporcionan herramientas para poner en contacto a los amigos de los amigos o para establecer contactos on-line.

2.7.4 Mashup

Es otra de la herramientas [33] más populares de la web 2.0, que significa “mezcla de aplicaciones” o “aplicaciones compuestas”. El mashup permite utilizar contenido procedente de diferentes páginas web para alimentar la página web propia. Por un lado están las API (*Application Programming Interface* o servicios web) propuestas por editores de contenido, y por otro, las herramientas para mezclar este contenido con los de la página del usuario (utilizando tecnologías Ajax o XML).

Uno de los ejemplos más conocidos que se beneficia de esta tecnología es, el del periódico New York Times, en su mashup se combinan por ejemplo, los datos cartográficos de una ciudad con la información sobre una huelga de transportes y los comentarios de los lectores sobre su impacto en el ámbito local.

2.8 LA WEB 2.0

La web 2.0 es un término que se ha usado varias veces, y aunque parezca el nombre de una tecnología o de un producto, no es nada nuevo, tan solo designa la evolución en el modo en el que la gente utiliza la red (**Figura 7**) [50].

El fundador de Amazon, Jeff Bezos, ofrece quizás la definición más acertada del objetivo de la web 2.0: “hacer útil Internet” [33].

La idea básica es que la aplicación original de la red, Web 1.0, funcionaba según un modelo en el que el editor ponía su información en una página web a disposición de los interesados, y los clientes utilizaban los navegadores para enviar una solicitud de información que les era entonces devuelta en una página cada vez, y si se quería más información o una actualización había que apretar un botón y esperar a que los datos fueran transferidos y la pantalla refrescada.

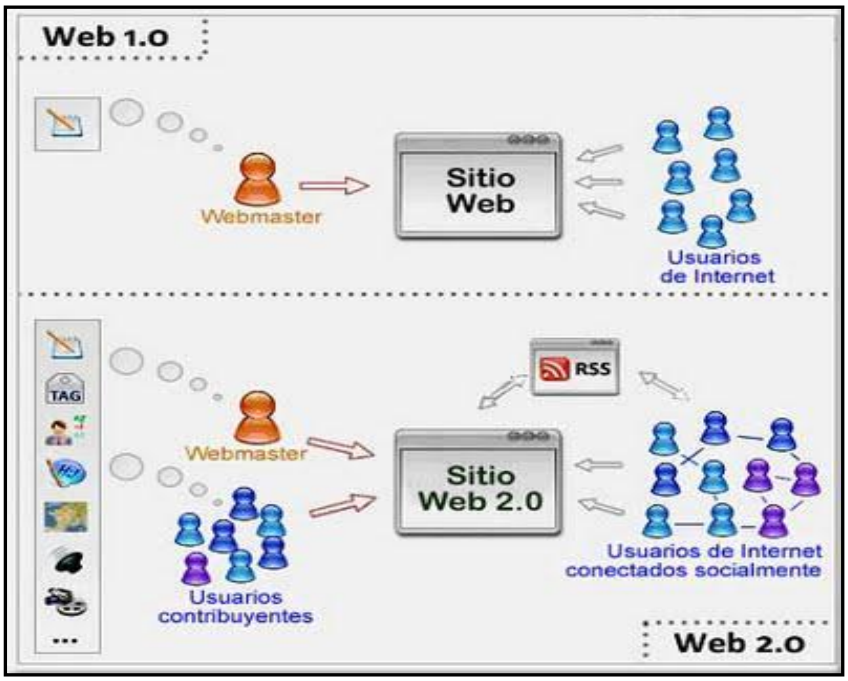


Figura 7 — Web 1.0 vs Web 2.0

Éste es el modo en el que todavía trabajan la mayoría de las webs, sin embargo, la web 2.0 difiere en tres aspectos fundamentales: el público se está convirtiendo en el editor de información, es él el que define cómo quiere ver la información y está constituyendo comunidades en este proceso.

Asimismo, dentro de esta evolución del uso de la red, se puede ver como la web 1.0 consistía en grandes empresas de software que lanzaban nuevas aplicaciones al mercado, mientras que la web 2.0 está más bien movida por el intento de crear una comunidad de desarrolladores que comparten códigos y aplicaciones personalizadas.

Algunos ejemplos que muestran como ahora el usuario es el que tiene el control son: Meebo, que permite chatear con amigos en cuatro sistemas de mensajería instantánea; Amazon Light, web desarrollada de forma independiente a Amazon, que modifica la interfaz de usuario pero con los mismos contenidos; o Rollyo [51], que permite crear un buscador personalizado que solo consulta las fuentes que se le especifiquen.

2.8.1 RSS

Además de publicar sus propios contenidos, los usuarios también deciden como quieren ver el contenido de otras fuentes. El RSS, siglas de *Really Simple Syndication* [52] (“servicio de noticias sencillo”) es una tecnología que permite la catalogación de información completamente adaptada a las preferencias de los usuarios.

Los archivos RSS contienen metadatos sobre fuentes de información especificadas por los internautas. Su función principal consiste en avisar automáticamente a los usuarios cuando los recursos que ellos han seleccionado sean actualizados, mostrando el cambio sin que aquéllos tengan que comprobar directamente las páginas. Se trata de un formato de transmisión de información basado en el lenguaje de etiquetado XML, que se utiliza para clasificar cada artículo por título, descripción y enlace directo.

Actualmente, los principales medios de comunicación de todo el mundo también presentan su contenido en RSS, y permiten subscribirse a su servicio de forma gratuita. De esta manera los consumidores de noticias ya no visitan CNN o El País en busca de titulares, en su lugar, se registran en diferentes hilos de información a través de un agregador como puede ser Bloglines (www.bloglines.com).

El RSS generalmente se considera como exclusivamente limitado al campo de la publicación y suscripción a titulares de noticias y alimentadores blog, pero es una tecnología innovadora que puede ser empleada en miles de aplicaciones posibles.

Para definir un código RSS [52], lo único que se debe hacer será definir el lenguaje de marcado y el tipo de caracteres que se va a utilizar, escoger una de las tres especificaciones de RSS que existen, y crear un "canal" en el que introducir los contenidos que se quieren mostrar a los demás usuarios (**Figura 8**).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rss version="2.0">
 <channel>
  <title>NETT PERU ... Conectando Negocios</title>
  <link>
        http://www.nettperu.com/
  </link>
  <description>Hosting (Alojamiento Web) desde S/.100 anuales</description>
  <item>
   <title>Acepte tarjetas de creditos en su pagina web</title>
   <link>http://nettperu.com/contacto.html</link>
   <description>Venda sus servicios y/o productos en Internet.</description>
  </item>
  <item>
   <title>Aprenda a enviar su publicidad por Internet</title>
   <link>http://www.fullpracticos.com/cursos/envio_publicidad/publicidad.html</link>
   <description>Curso practicos para crear y enviar su publicidad por Internet.</description>
  </item>
 </channel>
</rss>
```

Figura 8 — Ejemplo de código RSS

2.9 SISTEMAS WEB DE SEGUIMIENTO Y ANÁLISIS DE MEDIOS DE COMUNICACIÓN EN EL MERCADO

En Internet existen páginas dedicadas tanto a explotar la información de noticias en el tiempo, como al seguimiento de las mismas en diferentes medios de comunicación, algo similar al proyecto desarrollado, como son:

2.9.1 News Brief

Se trata de un sistema informático que selecciona y coloca los reportajes de forma automática cada 10 min, 24 horas al día (**Figura 9**) [53].

Todas las noticias están categorizadas, realizando un seguimiento en el tiempo (por horas) de las 10 noticias más importantes del día, indicando el número de artículos encontrados en distintos medios de comunicación. Pero además como complemento, el usuario puede subscribirse a su RSS, o descargarse un archivo .KML que muestra lugares e información relevante, pertenecientes a dichas noticias, en el sistema Google Earth.

Figura 9 — News Brief

2.9.2 Tus Titulares

Tus titulares es un portal que indexa noticias, y que las muestra ordenadas por fecha descendente de aparición en los medios de comunicación. Además, posee un buscador que permite agilizar la exploración del tema que se desea (**Figura 10**) [54].

Sin embargo, lo más relevante es la participación activa del usuario final, ya que puede elegir sus propios titulares para crearse su página principal de noticias, así como comentar cualquier artículo, convirtiendo a la página en una mezcla de portal de noticias y blog.

Figura 10 — Tus Titulares

2.9.3 Spy Press

Sistema informático que proporciona resúmenes de noticias personalizados para el usuario, cada día a partir de las 7:00AM, en la hora acordada, y en el formato acordado. Tiene una cobertura de medios españoles que incluye unos 2.500 medios de comunicación de internet y las principales cabeceras de prensa escrita de España [55].

Se trata de un sistema de pago, como muchos otros existentes en el mercado, que se especializa en dar servicios a empresas que precisan una información específica sobre algún sector en particular o necesitan conocer los puntos de vista de otros medios.

CAPÍTULO 3. TECNOLOGÍAS INVOLUCRADAS

3.1 INTRODUCCIÓN

En este capítulo se detallan los conceptos clave, que serán la base, para la comprensión del proyecto. Es necesario familiarizarse con las tecnologías utilizadas y los lenguajes usados en el desarrollo del proyecto, ya que actualmente existen multitud de alternativas en el mercado, que avanzan a un paso tan rápido que incluso para los que se dedican a utilizarlas les cuesta mantenerse al corriente de su aparición y utilidades.

En cada apartado se muestran las principales características de cada tecnología, estándar o conceptos usados durante el desarrollo del proyecto, así como algunas de las alternativas que existen en la actualidad y una justificación de su elección.

3.2 HTML, XML, XHTML

En primer lugar es importante conocer cómo crear una página web y que lenguajes son los más utilizados en la actualidad, dado que el sistema desarrollado será una aplicación web. A la hora de crear el Proyecto, se optó por el uso de HTML frente a XHTML, que combina sintaxis XML, debido a que este último es más estricto en cuanto a requisitos de código. No obstante se detallará de forma resumida los conceptos claves de ambos lenguajes, así como una introducción al XML, para facilitar la comprensión del lenguaje XHTML.

3.2.1 HTML

El estándar HTML (acrónimo de *Hypertext Markup Language*) fue desarrollado por el consorcio internacional W3C [57], fundado para el desarrollo de estándares Web, como lenguaje de marcado para el diseño y estructuración de textos para su posterior representación en forma de hipertexto, el formato estándar de las páginas Web.

HTML [58] permite publicar documentos en línea con encabezamientos, texto, tablas, imágenes, listas, etc. Acceder a documentos en línea cargándolos directamente en el navegador, o simplemente, haciendo clic en un hiperenlace existente en otro documento. Rellenar formularios con información y enviar esa información a un servidor que la procesará y realizará cierta tarea con ella. O incluir objetos multimedia en los documentos (como vídeos, imágenes, sonidos, animaciones) [12].

Es un lenguaje sencillo y editable desde cualquier editor de textos básico, además de con cualquier programa especializado para editar sitios web o código HTML.

La característica principal es la utilización de pares de etiquetas o marcas (de ahí que se conozca como lenguaje de marcado) de la forma <etiqueta-inicio></etiqueta-fin> para estructurar el texto. Estas etiquetas indicarán la forma en que se representará el texto, imágenes, tablas o demás elementos a representar en el navegador del dispositivo utilizado. A su vez, estas etiquetas tienen atributos para proporcionar al usuario una amplia variedad de características a utilizar a la hora de crear contenido para una página Web.

3.2.2 XML

XML [59] (*Extensible Markup Language*) es un lenguaje de etiquetado extensible muy simple pero estricto, desarrollado por el W3C como simplificación y adaptación de SGML, lenguaje para la organización y etiquetado de documentos.

XML es un metalenguaje similar a HTML cuya función principal es, en lugar de mostrar datos, describirlos. Además, no sólo es aplicable en Internet sino que también se propone como un estándar para el intercambio de información estructurada entre diferentes plataformas. Así se puede usar también en bases de datos, editores de texto y hojas de cálculo entre otros.

Las tecnologías XML son un conjunto de módulos que ofrecen servicios útiles a las demandas más frecuentes por parte de los usuarios. XML sirve para estructurar, almacenar e intercambiar información. Su importancia radica en que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

Por otro lado, los documentos XML deben estar "bien formados", es decir, cumplir con las definiciones básicas de formato y poder ser analizados correctamente por cualquier analizador sintáctico que cumpla con la norma.

3.2.3 XHTML

Ante la llegada de nuevos dispositivos surge XHTML (*Extensible Hypertext Markup Language*) [60] que combina la sintaxis de HTML para mostrar datos, con la sintaxis XML para describirlos.

XHTML refuerza la separación entre contenido y presentación. Se elimina así aquellos elementos y atributos relacionados con estilo (fuentes, colores, etc.). La ventaja es poder facilitar el cambio de la información de presentación para adaptarla a los distintos dispositivos de salida (teléfono móvil, televisor, ordenador, etc.).

Además es más estricto porque exige unos requisitos en cuanto al código que debe tener el documento como la estructuración coherente dentro del mismo: elementos correctamente anidados, etiquetas en minúsculas, elementos cerrados correctamente, etc.

A partir de la versión XHTML 1.1, se inició un proceso de modularización de la especificación. Esto permite que los dispositivos con capacidades reducidas o limitadas implementen únicamente aquellos módulos de la especificación que sean pertinentes.

3.3 CSS: estilo de los documentos

La evolución de la recomendación HTML sigue la línea de separar el contenido de la presentación, introduciendo mecanismos que permitan representar la información de estilo. Surgen así las hojas de estilo, *Cascade Style Sheet* (CSS) [12] y [61].

El CSS se usa para dar estilo a los documentos HTML, XML y XHTML. Con su utilización se pretende separar el contenido de los documentos de su presentación, permitiendo a los desarrolladores Web elegir el formato de múltiples páginas al mismo tiempo. Funciona mediante reglas que son declaraciones de estilo sobre uno o más elementos. De esta manera, una hoja de estilo se compone de una o varias de estas reglas aplicadas sobre un documento.

3.4 Páginas Web dinámicas

El contenido de una página puede ser predeterminado ("página web estática") o generado al momento de visualizarla o solicitarla a un servidor web ("página web dinámica"). Las páginas dinámicas que se generan al momento de la visualización se hacen a través de lenguajes interpretados, generalmente JavaScript, y la aplicación encargada de visualizar el contenido es la que debe generarlo.

Mientras que una página Web estática está realizada íntegramente en XHTML o HTML y se caracteriza por la ausencia de movimiento y funcionalidades, y una absoluta opacidad a los deseos o búsquedas del visitante, una página Web dinámica utiliza diversos lenguajes y técnicas de programación, permitiendo un gran número de posibilidades en su diseño y desarrollo, además de un proceso de actualización sumamente sencillo, sin necesidad de entrar en el servidor.

Dado que el sistema de seguimiento y análisis desarrollado no solo muestra información estática (HTML), si no que requiere una interacción por parte del usuario para

mostrar los contenidos, necesitaremos del uso de lenguajes interpretados que permitan integrarse con código HTML y que ofrezcan el dinamismo deseado.

Aunque existen multitud de lenguajes en el mercado que podrían haberse usado para el desarrollo del proyecto: PHP, ASP.Net, Jsp, Servlets, Phython o Perl, entre otros a continuación se definirán algunos de los más extendidos por su uso y cuál ha sido la decisión final seleccionada.

3.5 Lenguajes de programación

3.5.1 JavaScript

JavaScript [62] es un lenguaje de programación creado por Netscape con el objetivo de integrarse en HTML y facilitar la creación de páginas dinámicas, sin necesidad de utilizar scripts de CGI, siendo implementado en todo tipo de navegadores web [22].

No hay que confundir Java con JavaScript. Java es un lenguaje completo que permite crear aplicaciones independientes, mientras que JavaScript es un lenguaje que funciona como extensión del HTML. Es un lenguaje de programación orientado a objetos, diseñado para el desarrollo de aplicaciones cliente-servidor a través de Internet.

El código de programa de JavaScript, llamado script, se introduce directamente en el documento HTML y no necesita ser compilado, es el propio navegador el que se encarga de traducir dicho código.

3.5.2 CGI

El CGI [63] (*Common Gateway Interface*), fue la primera manera práctica de crear contenido dinámico para las páginas web, proporcionó el mecanismo original a través del cual los usuarios podían realmente ejecutar programas en servidores Web y no simplemente pedir páginas HTML. Sin embargo, cuando la Red incrementó su popularidad y las demandas de

tráfico situadas en sitios Web aumentaron, CGI no era lo suficientemente eficiente para continuar. Esto es debido a que con CGI, cada vez que se recibe una petición, el servidor tiene que empezar a ejecutar una nueva copia del programa externo, lo que resulta ineficiente.

3.5.3 PHP

PHP (*Hypertext Preprocessor*), de acuerdo con el sitio oficial de PHP, que puede encontrarse en *www.php.net*, es un lenguaje interpretado de alto nivel embebido en páginas HTML y ejecutado en el servidor Web. La mayor parte de su sintaxis ha sido tomada de C, Java y Perl con algunas características específicas de sí mismo. La meta del lenguaje es permitir a los desarrolladores la generación dinámica de páginas rápidamente.

PHP al ser un lenguaje que se ejecuta en el servidor no es necesario que un navegador lo soporte, es decir, es independiente del navegador, pero sin embargo para que una página PHP funcione el servidor donde esté alojada debe soportar PHP (**Figura 11**).

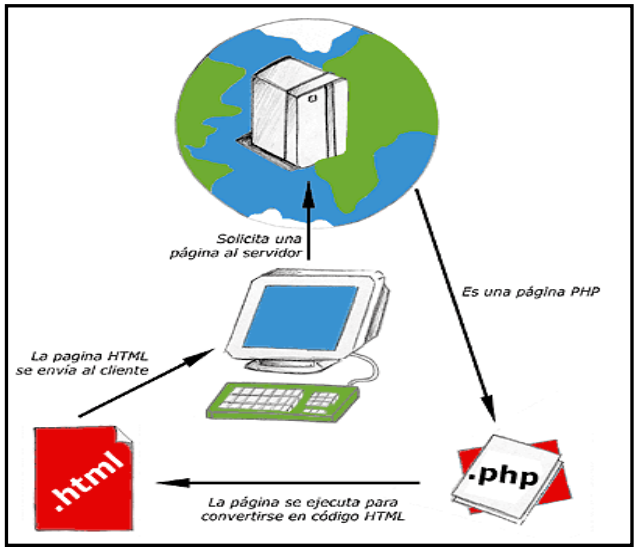


Figura 11 — Petición de una página en PHP

Una de sus características más importante, y uno de los motivo de su elección en el proyecto, es su soporte para gran cantidad de bases de datos como InterBase, MySQL, y Oracle, entre otras.

Debido a su amplia distribución, PHP está perfectamente soportado por una gran comunidad de desarrolladores. Como producto de código abierto, PHP goza de la ayuda de un gran grupo de programadores, permitiendo que los fallos de funcionamiento se encuentren y se reparan rápidamente [28].

3.5.4 ASP.NET

ASP (*Active Server Pages*) [67] es una tecnología de Microsoft para la construcción de sitios web dinámicos y conducidos por bases de datos, haciendo que resulte muy sencillo crear las aplicaciones web dinámicas que podemos encontrar hoy en día por todo Internet.

Proporciona un modelo simple de programación basado en el Framework .NET, y varios controles y servicios ya incluidos que permiten crear los escenarios que encontramos en gran parte de las aplicaciones, con poco código y esfuerzo.

ASP.NET representa una separación radical de versiones anteriores de ASP, por ejemplo utiliza código compilado escrito en *Common Language Runtime* como Visual Basic y C, en lugar de lenguajes script interpretados como VBScript. Además, incluye mecanismos de almacenamiento de datos y de páginas en la memoria temporal, lo que permite mejorar el rendimiento de un sitio web. Las páginas consisten en código y marcas que son compiladas y ejecutadas dinámicamente en el servidor para producir una traducción para el navegador.

En general, sus características son muy similares a las de PHP, sin embargo, ASP.NET no es de libre distribución, es más complejo de aprender y usar, y por las múltiples funciones que tiene es un poco más lento que PHP.

3.5.5 JSP Y SERVLETS

Las páginas de servidor Java, o JSP [66], es una tecnología utilizada para crear sitios Web dinámicos o interactivos que se basa en el lenguaje de programación Java, con todas las prestaciones de que dispone [9].

Las JSPs se construyen sobre una tecnología Java anterior para crear contenido Web dinámico, los servlets de Java. De hecho las JSP son en realidad servlets, ya que una página JSP se compila a un programa en Java la primera vez que se invoca, y del programa en Java se crea una clase que se empieza a ejecutar en el servidor como un servlet.

La principal diferencia entre los servlets y las JSP es el enfoque de la programación: un JSP es una página Web con etiquetas especiales y código Java incrustado, mientras que un servlet es un programa que recibe peticiones y genera a partir de ellas una página web.

Los servlets (un “mini-servidor”) son otra tecnología alternativa a CGI. Los servlets, por ellos mismos, no son aplicaciones autónomas; se cargan en la memoria mediante un contenedor de servlet. Este contenedor de servlet funciona entonces como un servidor Web, recibiendo peticiones HTTP de los navegadores Web y pasándolas a los servlets.

Por otro lado, es importante comprender la distinción entre JavaScript y Páginas de Servidor Java, ya que, mientras que el código JavaScript se ejecuta generalmente por el cliente Web (navegador) después de que el servidor Web haya enviado la respuesta HTTP al navegador, las Páginas de servidor Java son ejecutadas por el servidor Web antes de que éste envíe la respuesta HTTP. Así, se dice que JavaScript es una tecnología aplicada al cliente y su código subyacente puede ser visto (y copiado) por usuarios Web, mientras que las JSP son una tecnología del lado del servidor y su código subyacente no está expuesto a los usuarios Web; es procesado por el servidor Web antes de que llegue al cliente.

3.5.6 Elección de PHP frente a otros lenguajes

De un tiempo a esta parte, se ha venido observando un ascenso imparable en la utilización de PHP para la creación de páginas web dinámicas, esto es debido a que cada vez son más los programadores que se apoyan en este lenguaje para el funcionamiento de sus aplicaciones.

Al diseñar sitios web, las principales alternativas a PHP, como se ha detallado, son: simplemente HTML, los scripts CGI, ASP, JSP y Perl o JavaScript aunque este no es realmente una alternativa a PHP, ya que es una tecnología del lado del cliente y no puede utilizarse para crear páginas HTML por sí mismo.

En base a los detalles descritos y algunas ventajas que sobresalen por encima del resto de lenguajes se decidió el uso de PHP junto con JavaScript y todo ello embebido en HTML para la creación del sistema desarrollado.

Entre las ventajas de PHP frente al resto de lenguajes de scripting en el lado del servidor se puede destacar:

- Software libre y de código fuente abierto.
- Rápido de programar y diseñado para la Web.
- Plataforma cruzada: puede utilizarse en prácticamente todos los sistemas operativos del lado del servidor.
- Soporte nativo para prácticamente cualquier Base de Datos.
- Perfecta integración con Apache y MySQL.
- Sintaxis clara y bien definida.
- Sencillo de aprender y además cuenta con una buena documentación.
- Modulable.

3.6 BASES DE DATOS

Todo motor de búsqueda necesita de una base de datos y un programa de indización para gestionarla. Durante la creación del proyecto se bajaron varias hipótesis para este fin, en primer lugar se barajó el uso de sistemas basados en texto que seleccionaran aquellos documentos que nos fueran de utilidad como es Lucene o Xapian, pero tras el estudio de estas tecnologías que serán descritas más adelante, se optó mejor por usar un sistema de bases de datos relacional, concretamente MySQL, que simplificaba mucho el trabajo a realizar.

En el mercado existen multitud de sistemas de administración de bases de datos gratuitos, como son MySQL, PostgreSQL o SQLite, y otros no libres como serían Oracle o Microsoft SQL Server. Pero además también existen estándares que permiten la ejecución de operaciones sobre bases de datos como son la API JDBC, desarrollado por Java, u ODBC, desarrollado por Microsoft.

Dado la extensión que sería hablar de todas estas tecnologías se ha optado por hacer una comparativa entre MySQL y PostgreSQL, dado que finalmente se eligió MySQL como sistema de administración, y además hacer una breve introducción a uno de los estándares de acceso a bases de datos como es JDBC para entender cuál es su funcionamiento.

3.6.1 Teoría de bases de datos

Una base de datos es una colección de archivos interrelacionados, creados y mantenidos mediante un sistema de gestión de bases de datos o SGBD (en inglés *database management system*, abreviado DBMS), que es un tipo de software muy específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizadas.

Una base de datos se diseña, construye y rellena con datos para un propósito específico. Dispondrá de un conjunto de usuarios interesados en alguna de las aplicaciones en ella creadas.

El contenido de una base de datos engloba la información (almacenada en archivos) de una organización, de tal manera que los datos estén disponibles para los usuarios, y así intentar minimizar la redundancia.

3.6.2 SQL

El nombre SQL significa Lenguaje de Consulta Estructurado (*Structured Query Language*). Fue diseñado e implementado por IBM Research a modo de interfaz para un sistema de bases de datos relacional conocido como SYSTEM R [8].

Una base de datos relacional se puede definir simplemente como una base de datos conformada por tablas y columnas que se relacionan entre sí. Estas relaciones están basadas en valores clave contenidos en una columna.

SQL es un lenguaje de bases de datos global, cuenta con sentencias para definir datos, consultas y actualizaciones. También dispone de características para definir vistas en la base de datos, especificar temas de seguridad y autorización, definir restricciones de integridad, y especificar controles de transacciones.

Los tipos de datos básicos disponibles para los atributos de la base de datos son: numérico, cadena de caracteres, cadena de bits, booleano, fecha y hora.

3.6.3 MySQL

MySQL es un sistema de administración de bases de datos relacionales Open Source (código abierto) y licencia pública, que opera a través de un intérprete de comandos. Se trata de un servidor SQL diseñado para grandes cargas y que puede procesar consultas muy complejas [19].

Como sistema de bases de datos relacional, MySQL permite combinar multitud de tablas diferentes para optimizar la eficacia y la velocidad. Además es un servidor

multiprocesos, lo que significa que cada vez que alguien establece una conexión con el servidor, el programa servidor crea un subproceso para manejar la solicitud del cliente. Esto hace al servidor extremadamente rápido.

Otra característica muy valiosa de MySQL es su sistema de ayuda en línea. Todos los comandos de MySQL se introducen mediante el indicador de comandos. Además, MySQL ha sido portado a casi cualquier plataforma, con lo que no es necesario cambiar de plataforma para poder usarlo [14].

3.6.4 PostgreSQL

PostgreSQL [69] es un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el sistema de gestión de bases de datos de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales.

PostgreSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando.

Sus características técnicas la hacen, una de las bases de datos más potentes y robustas del mercado.

Su desarrollo comenzó hace más de 15 años, y durante este tiempo, estabilidad, potencia, robustez, facilidad de administración e implementación de estándares han sido las características que más se han tenido en cuenta durante su desarrollo. PostgreSQL funciona muy bien con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema.

Sin embargo, PostgreSQL, consume más recursos y carga del sistema que MySQL, lo que causa que sea más lento. Además en cuanto a consideraciones de estabilidad del servidor, en general, parece que MySQL es más estable, y finalmente MySQL está mejor integrado junto con PHP y Apache.

3.6.4 JDBC

JDBC (*Java Database Connectivity*) [68] es una API para la ejecución de sentencias SQL, que permite la realización de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del software de administración o de la base de datos a la cual se acceda. Y además proporciona una base común para la construcción de herramientas y utilidades de alto nivel [20].

Una API (*Application Programming Interface*), no es más que una interfaz de programación de aplicaciones, provista por los creadores del lenguaje Java, y que da a los programadores los medios para desarrollar aplicaciones Java.

Como el lenguaje Java es un lenguaje orientado a objetos, la API de Java provee de un conjunto de clases utilitarias para efectuar toda clase de tareas necesarias dentro de un programa.

La API JDBC hace que sea muy fácil escribir código necesario para acceder a bases de datos relacionales ya que ofrece un amplio conjunto de clases e interfaces de Java que encapsulan las funcionalidades necesarias.

Para usar JDBC con un sistema gestor de base de datos en particular, es necesario disponer del driver JDBC apropiado que haga de intermediario entre ésta y JDBC. Dependiendo de varios factores, este driver puede estar escrito en Java puro, o ser una mezcla de Java y métodos nativos JNI (*Java Native Interface*).

3.6.6 Elección de MySQL como sistema de administración de bases de datos

Como se ha comentado anteriormente, en el mercado existen varios sistemas libres o gratuitos entre los que poder elegir: MySQL, Apache Derby, PostgreSQL, o algunos de los motores de base de datos gratuitos de las marcas comerciales pero que no disponen de mantenimiento. Cuando se compara MySQL con otros sistemas de bases de datos, hay muchos aspectos que sopesar como son: rendimiento, mantenimiento, características (conformidad SQL, extensiones, etc.), o incluso las condiciones y restricciones de la licencia.

Teniendo en cuenta estas consideraciones, MySQL tiene muchas características atractivas que ofrecer [7]:

- **Velocidad:** MySQL es rápido. Los desarrolladores sostienen que MySQL es posiblemente la base de datos más rápida que se pueda encontrar.
- **Facilidad de uso:** MySQL es un sistema de base de datos de alto rendimiento pero relativamente simple y es mucho menos complejo de configurar y administrar que sistemas más grandes.
- **Coste:** MySQL es gratuito para la mayoría de usos internos.
- **Capacidad de gestión de lenguajes de consulta:** MySQL comprende SQL, el lenguaje elegido para todos los sistemas de base de datos modernos, aunque también se puede acceder empleando aplicaciones que admitan ODBC, protocolo de comunicación de bases de datos desarrollado por Microsoft.
- **Capacidad:** pueden conectarse muchos clientes simultáneamente al servidor.
- **Conectividad y seguridad:** MySQL está completamente preparado para el trabajo en red y las bases de datos pueden ser accedidas desde cualquier lugar en Internet.
- **Portabilidad:** funcionamiento entre plataformas.

Además la comunidad MySQL, desarrolladores y no desarrolladores por igual, son muy participativos, proporcionando, a los usuarios que quieran subscribirse, recursos de mantenimiento, ayuda en línea, etc.

3.7 RECUPERACIÓN DE INFORMACIÓN. SISTEMAS BASADOS EN TEXTO

Como ya se conoce, la búsqueda y recuperación de información, consiste en una serie de procesos donde se accede a una información previamente almacenada, mediante herramientas informáticas que permiten establecer ecuaciones de búsqueda específicas. Para ello, dicha información ha debido de ser estructurada previamente a su almacenamiento.

Entre las herramientas de recuperación se tienen: bases de datos, sistemas de indización, que extraen uno o más conceptos que representan el contenido temático del documento con el objetivo de recuperarlo posteriormente, sistemas basados en texto, etc. En el desarrollo del proyecto se ha utilizado como sistema de recuperación de información, para crear el buscador, un sistema de bases de datos relacional, como es MySQL.

Sin embargo, también es importante destacar otros tipos de sistemas de recuperación de información de manera automática, como son los sistemas basados en texto, que seleccionan aquellos textos o documentos que son adecuados a una necesidad del usuario entre un conjunto más amplio.

A continuación se describen algunos de los sistemas de recuperación de información de código libre existentes.

3.7.1 Swish-E

Swish-e (**Figura 12**) (*Simple Web Indexing System for Humans – Enhanced*) [70] es un sistema gratuito, rápido y flexible que permite la indexación de colecciones de páginas web u otros archivos, que puede implementarse fácilmente en una Intranet o portal corporativo.

Swish-e está diseñado para colecciones de hasta un millón de documentos y utiliza un modelo de recuperación de información booleano.

Mediante las instrucciones de un sencillo archivo de configuración, Swish-e recorre los directorios y archivos, y genera un índice que puede ser utilizado en cualquier plataforma soportada. A partir del índice, se pueden realizar búsquedas desde la línea de comandos, a través de una librería C y también por medio de una interfaz web implementada en un script en Perl.

Figura 12 — Logo Swish-e

Swish-e puede indexar texto plano, e-mail, PDF, HTML, XML, Microsoft Word, Power Point, Excel, y casi cualquier archivo que pueda convertirse a XML o HTML. También se usa como suplemento de bases de datos como MySQL para búsqueda rápida de texto.

3.7.2 Xapian

Xapian [71] (**Figura 13**) es un proyecto basado en el código *Open Muscat* (también conocido como Omsee) originalmente desarrollado por *BrightStation*. *BrightStation* abandonó el proyecto, pero algunos de los desarrolladores originales y otros de la comunidad *Open Muscat* decidieron continuar. Un número creciente de organizaciones y proyectos usan Xapian entre ellos se incluyen Orange, Gmane Y Die Zeit [6].

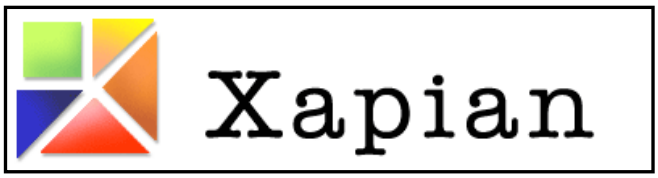


Figura 13 — Logo Xapian

Xapian está diseñado como una herramienta fácilmente adaptable para permitir a los desarrolladores la posibilidad de añadir a sus propias aplicaciones de una manera sencilla las capacidades de búsqueda e indexación. Xapian implementa el modelo probabilístico de obtención de información y entre sus características destaca, que es un software libre, que permite el uso de Unicode, y además almacena datos indexados en UTF-8, que es un sistema multiplataforma y está escrito en C++, que realiza transacciones, lo que garantiza que si la base de datos falla en mitad de la actuación se mantendrá en un estado consciente, o por ejemplo que permite la búsqueda de términos de una misma familia de palabras (en múltiples idiomas), lo que permite encontrar documentos relevantes que de otra manera serían descartados.

Puede indexar HTML, PHP, PDF, PostScript, OpenOffice/StarOffice, OpenDocument, Microsoft Word, Excel, PowerPoint, RTF, documentación Perl POD, pero también permite indexación de otros formatos usando filtros conversores. Y además también permite indexar datos de SQL, u otros RDBMS, lo que incluye MySQL, PostgreSQL, SQLite, Oracle, DB2, MS SQL, LDAP y ODBC.

3.7.2 Lucene

Lucene [72] (**Figura 14**) es una herramienta que permite tanto la indexación cómo la búsqueda de documentos. Creada bajo una metodología orientada a objetos e implementada completamente en Java, no se trata de una aplicación que pueda ser descargada, instalada y ejecutada sino de una API flexible, muy potente y realmente fácil de utilizar, a través de la cual se pueden añadir, con pocos esfuerzos de programación, capacidades de indexación y búsqueda a cualquier sistema que se esté desarrollando.

Originalmente fue implementado en Java por Doug Cutting, pero posteriormente aparecieron versiones en lenguajes como Perl, C++, Python y PHP [6].

Está financiado por la fundación Apache Software y algunos de sus proyectos tales como Nutch y Solr, están basado en Lucene.

Esta librería se adapta a cualquier aplicación que requiera indexación y búsqueda completas de texto. Lucene es ampliamente conocido por su utilidad en la implementación de motores de búsqueda en Internet y locales.

Figura 14 — Logo Lucene

El principio fundamental de la arquitectura lógica de Lucene consiste en un documento que contiene campos de texto. Textos de documentos como PDF, HTML, Microsoft Word y otros muchos pueden ser indexados siempre que de ellos se pueda extraer su información textual.

Entre las características que ofrece Lucene a través de su API se pueden destacar:

- Alto rendimiento: escasos requerimientos de RAM, indexación incremental y tamaño de la indexación de entre el 20 y 30% del tamaño del texto indexado.
- Algoritmos de búsqueda, eficientes, precisos y potentes, permite realizar búsquedas clasificadas, distintos tipos de peticiones, búsqueda por campos o por fecha, clasificación en base a cualquier campo, combinación de resultados de búsquedas multi-índice, permite actualización y búsqueda simultáneas.
- Independiente de la plataforma, se trata de un software libre bajo la licencia de Apache que permite su uso en cualquier plataforma.
- Las clases claves para construir el motor de búsqueda son:
  - ***Document***: representa un documento en Lucene. Se indexa un objeto documento y se obtiene un objeto documentado cuando se realiza una búsqueda.
  - ***Field***: representa una sección de un documento (*Document*). Este objeto contendrá un nombre para la sección del dato actual.

- ***Analyzer***: es una clase abstracta que se usa para proporcionar una interfaz que permitirá que un documento pueda ser indexado.
- ***IndexWriter***: clase usada para crear y mantener índices.
- ***IndexSearcher***: clase usada para buscar en un índice.
- ***QueryParser***: clase que se usa para poder crear criterios para buscar a través de un índice.
- ***Query***: clase abstracta que contiene los criterios de búsqueda creados por *QueryParser*.

CAPÍTULO 4. DESARROLLO DEL PROYECTO

4.1 ARQUITECTURA DEL SISTEMA

A la hora de diseñar un sistema de seguimiento de medios de comunicación en Internet, el primer problema que se plantea es la elección entre crear un directorio o crear un motor de búsqueda. Como ya se sabe, la diferencia principal entre un motor de búsqueda y un directorio, es que el motor usa sistemas automáticos para indexar las páginas (robot o araña), mientras que los directorios usan personas para calificar cada página a indexar.

Teniendo en cuenta esta característica y en base a los requisitos del sistema a desarrollar, almacenar información diaria sobre las noticias publicadas en diferentes medios de ámbito nacional, se tomó la decisión de crear un motor de búsqueda, en lugar de un directorio que requiere más soporte humano y mantenimiento.

La arquitectura (**Figura 15**) del motor de búsqueda diseñado constará de las siguientes partes:

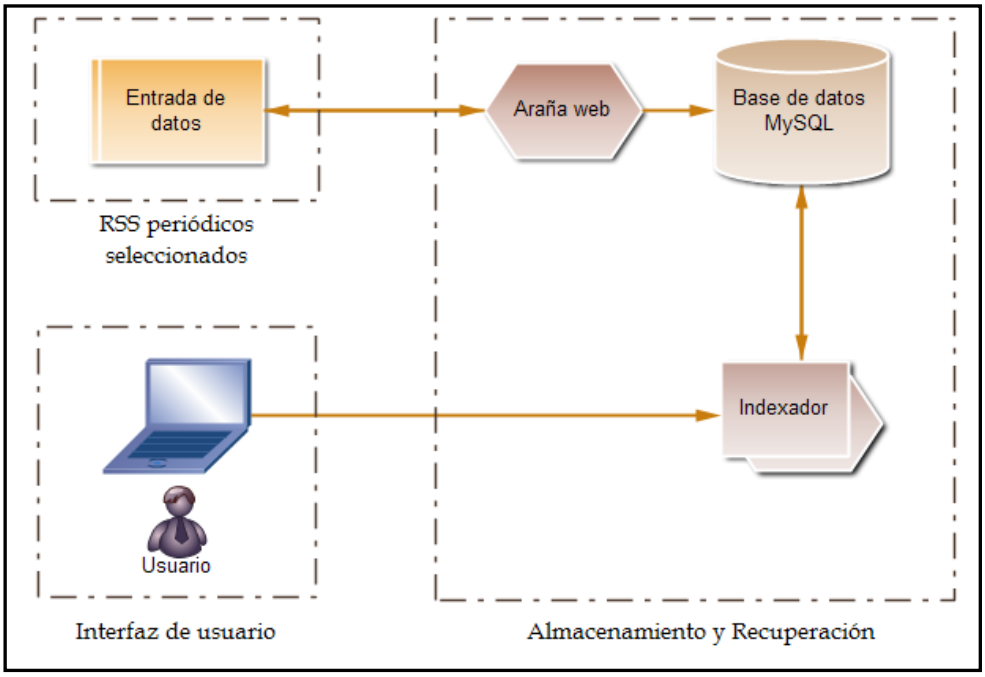


Figura 15 — Arquitectura del motor de búsqueda

1. **La araña**: como ya se ha explicado, la araña o crawler consiste en un programa que parte con una lista inicial de URLs, y se encarga de recopilar la información que interesa y almacenarla en la base de datos. El proyecto se centrará en la información proporcionada por distintos medios de comunicación en Internet, que aparte de mostrar sus contenidos de manera tradicional a través de sus distintas páginas HTML/XHTML, también presentan su contenido en RSS, un formato XML para compartir fácilmente la información.

El programa se encargará de inspeccionar una lista inicial de direcciones web en formato RSS, proporcionada de forma metódica y automatizada, y analizar la información que interesa almacenándola en la base de datos.

Para poder recorrer la lista de RSS se empleará SimplePie, una biblioteca de PHP que permite leer rápidamente y procesar canales (feeds) del tipo RSS o Atom [65].

SimplePie, tiene todos los métodos necesarios para obtener todos los datos del feed, incluso permite hacer caché para acelerar las consultas. Todo ello unificado en un solo objeto de forma que, para el usuario le resulte igual leer RSS o Atom (en cualquiera de sus versiones) sin cambiar ninguna línea de código.

Además es importante conocer que un feed es un archivo generado por algunos sitios web que contiene una versión específica de la información publicada en esa web, en jerga informática suele referirse a un tipo de dato empleado para suministrar información que es actualizada con frecuencia. Se emplea para denominar a los documentos con formato RSS o Atom, basados en XML, que permiten a los agregadores recoger información de páginas web sindicadas. Sin embargo, a menudo el término RSS (Really Simple Syndication) se usa erróneamente para referirse a una fuente web, independientemente de que el formato de dicha fuente web sea RSS o Atom.

2. **Base de datos**: constituida principalmente por una lista de elementos indizados, extraídos de la información que contienen los documentos RSS que generan algunas páginas web como alternativa a la información típica HTML/XHTML.

La base de datos diseñada contendrá una sola tabla que almacenará el título, el contenido, la URL y la fecha de publicación, de cada una de las noticias almacenadas.

Para gestionar la base de datos se empleará MySQL, un sistema de gestión de bases de datos relacional.

2. **Motor de indexación**: en el momento en que cualquier usuario hace una búsqueda en el motor, se inspecciona la base de datos, observando y comparando las palabras que el usuario escribió con cada registro que está guardado, sin embargo el algoritmo debe ser cuidadoso para no traer resultados basura y además tiene que hacer un filtrado para obtener lo que realmente desea el usuario, finalmente devolverá los resultados tomando en cuenta todos los elementos relevantes con las palabras que fueron especificadas inicialmente.

3. **Interfaz de recuperación web**: página web que contendrá el formulario de búsqueda, y que se encargará de emplear el motor de indexación para evaluar dicha búsqueda y hacerla coincidir con los documentos más relevantes en la base de datos, mostrando sólo los resultados adecuados.

Para el desarrollo de la interfaz se han seguido los criterios definidos por la disciplina de la interacción persona-computador, o IPO, más conocida por sus siglas en inglés, HCI (*Human Computer Interaction*), que estudia el intercambio de información entre las personas y los ordenadores. Su objetivo es que este intercambio sea lo más eficiente posible: minimizar los errores, incrementar la satisfacción del usuario, disminuir la frustración y en definitiva, hacer más productivas las tareas que relacionan a las personas y los ordenadores [73].

A la hora de escoger un buscador, los usuarios de éste no sólo esperan obtener buenos resultados (aunque sí es muy importante), sino que también valoran la interfaz que presente. Por ello, es recomendable seguir una serie de aspectos, cumpliendo en la medida de lo posible los principios de HCI, algunos puntos destacables a tener en cuenta son:

- La cantidad de opciones propuestas en la página principal debe tener un grado de complejidad tal que permita que el usuario pueda **aprender a utilizar el sistema en forma progresiva.**
- Suele valorarse una **interfaz simple**, en detrimento de aquellas que presentan además anuncios u otras ofertas. Tampoco suelen valorarse las interfaces en que se presentan muchos elementos en forma de opciones que son difíciles de localizar, o bien aquellas de diseño recargado. Estos aspectos pueden restar credibilidad al buscador.
- **Curva de Aprendizaje**, o cómo el usuario aprende a usar un buscador a medida que lo emplea. El aprendizaje de un producto y su usabilidad no son mutuamente excluyentes. El ideal es que la curva de aprendizaje sea nula, y que el usuario principiante pueda alcanzar el dominio total de la aplicación sin esfuerzo. Que el aprendizaje en el uso haya de ser sencillo no quiere decir que no se estudie cómo hacer el buscador lo más eficiente posible y usable también para usuarios expertos.
- En cuanto a la **posición de la herramienta para realizar una búsqueda**, resulta interesante que resulte accesible siempre. Es decir, que incluso mientras estemos consultando los resultados podamos encontrarla fácilmente, pues se presente en pantalla en un lugar visible.
- Los **resultados de la búsqueda** deben ocupar una parte destacada en la página, normalmente la central.
- En cuanto al **número de resultados obtenidos**, debe ser visible una vez realizada la búsqueda.

- Se deberá ofrecer una forma para visualizar posibles **más resultados de búsqueda**, más allá de los que aparecen en la primera página. Es recomendable que esta opción se ofrezca en la parte inferior de la página. Esto se debe a que, normalmente, el usuario deseará acceder a estas opciones una vez ha consultado los resultados que se le han ofrecido.
- El **contenido** debe ser **estándar**: enlaces en color azul oscuro y subrayados (la filosofía del azul para los enlaces), visitados en un tono rojo...
- **Iconos representativos y fácilmente localizables**. Presenta una especial importancia que resulte visible el botón a pulsar para proceder a realizar la búsqueda y que además sea identificable como tal.
- **Una sola aplicación o servicio**: la interfaz de usuario presenta la idea de la aplicación o servicio utilizado, en este caso el buscador, como un componente único. Es decir, la funcionalidad principal que debe ofrecerse es la realización de búsquedas. Otras que puedan ofrecerse son secundarias, y no debería darse les tanta relevancia en la página.
- Ley de Fitt, El tiempo para alcanzar un objetivo es una función de la distancia y tamaño del objetivo. Es por ello que es conveniente usar objetos grandes para las funciones importantes. En nuestro caso, el botón o el espacio de búsqueda.
- Por defecto al pulsar la **tecla "enter"** tras indicar una búsqueda, debería realizarse la misma y presentarse los resultados.

Finalmente el sistema de seguimiento de prensa diseñado vendrá acompañado de:

- **Función de análisis de noticias:** gráfico estadístico mediante el cual los usuarios podrán conocer las menciones que realizan los medios de comunicación sobre las noticias en las que estén interesados. Mostrará el total de apariciones de dicha noticia por día, a lo largo de un periodo de tiempo seleccionado por el usuario.

4.2 IMPLEMENTACIÓN DEL SISTEMA

De acuerdo a lo explicado en el apartado anterior, donde se describen los módulos que componen un motor de búsqueda, se puede dividir el desarrollo de este sistema en tres etapas básicas: creación de la araña o spider, integración y manipulación de la información en la base de datos y diseño de la interfaz de recuperación web y función de análisis.

4.2.1 Araña web

Los archivos RSS (**Figura 16**) se utilizan para crear y distribuir noticias, estos archivos contienen en su código XML un resumen de la información que contiene la página web, como se describió en apartados anteriores y que incluyen entre otros datos: enlaces, titulares y resúmenes de cada noticia, si bien, no toda la información mostrada en los RSS seleccionados será importante para el desarrollo del proyecto.

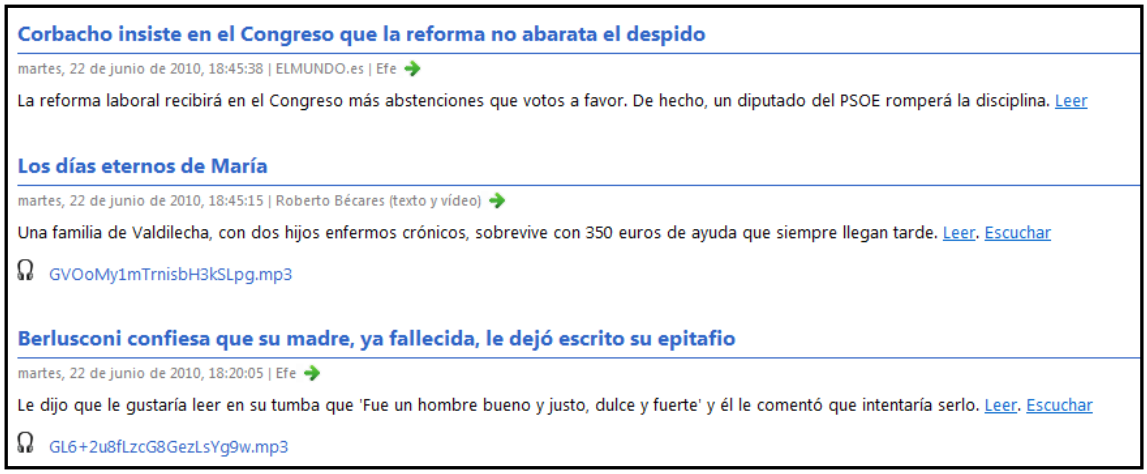


Figura 16 — Presentación web de la información en RSS

El código RSS de cada medio de comunicación contendrá en primer lugar una cabecera común (**Figura 17**) a todas las noticias con información relevante sobre la categoría de las noticias, o el titulo y la descripción del conjunto a mostrar. Esta información no será relevante para el sistema desarrollado y por tanto no será indexada.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:itunes="http://www.itunes.
<channel>
<title>Portada // elmundo.es</title>
<link>http://www.elmundo.es</link>
<description>Portada // elmundo.es</description>

<itunes:category text="News &amp;amp; Politics" />

<itunes:explicit>no</itunes:explicit>
<itunes:author>elmundo.es</itunes:author>
<itunes:image href="http://estaticos.elmundo.es/imagen/canalima.gif" />
<itunes:owner>
<itunes:email>podcasting@el-mundo.net</itunes:email>
<itunes:name>Unidad Editorial Internet S.L.</itunes:name>
</itunes:owner>
<language>es</language>
<copyright>(c) 2010, Unidad Editorial Internet, S.L.</copyright>
<pubDate>Tue, 12 Oct 2010 12:10:15 +0200</pubDate>
<lastBuildDate>Tue, 12 Oct 2010 12:10:15 +0200</lastBuildDate>

<category><![CDATA[News & Politics]]></category>

<ttl>60</ttl>
<atom:link href="http://estaticos.elmundo.es/elmundo/rss/portada.xml" rel="self" type="applicat
<image>
<title>Portada // elmundo.es</title>
<url>http://estaticos.elmundo.es/imagen/canalima144.gif</url>
<link>http://www.elmundo.es</link>
<width>144</width>
<height>24</height>
<description>elmundo.es</description>
</image>
```

Figura 17 — Código con la cabecera de información en cada RSS

A continuación por cada noticia publicada en el RSS <*item*> (**Figura 18**), se seleccionará su titulo <*title*>, enlace de redirección a la página principal donde se publicó el contenido <*link*>, el resumen de la noticia <*description*>, y la fecha de publicación <*pubDate*>, desechando información como el autor <*creator*>, imágenes asociadas o archivos multimedia <*media*:*content*> o información editorial <*guid*>.

```
<item>
<title><![CDATA[Fallece Manuel Alexandre, el secundario de oro del cine español]]></title>
<link>http://www.elmundo.es/elmundo/2010/10/12/cultura/1286866886.html</link>
<description><![CDATA[El intérprete de películas como &quot;Bienvenido, Mr. Marshall&quot; o &quot;Amanece
<dc:creator><![CDATA[Efe]]></dc:creator>
<media:description type="html"><![CDATA[El intérprete de películas como &quot;Bienvenido, Mr. Marshall&quot
<media:title type="html"><![CDATA[La capilla ardiente se instalará mañana en el Teatro Español]]></media:ti

<media:content url="http://estaticos.elmundo.es/elmundo/imagenes/2010/10/12/cultura/1286866886_4.jpg" mediu
<media:thumbnail url="http://estaticos.elmundo.es/movil/elmundo/imagenes/2010/10/12/cultura/1286866886_4_10
<itunes:duration>3:44</itunes:duration>
<itunes:author>elmundo.es</itunes:author>
<enclosure url="http://estaticos.elmundo.es/tts/rosa/06/24dNMyZKEEu3UomZS4GjcQ.mp3" length="739408" type="a
<guid>http://www.elmundo.es/elmundo/2010/10/12/cultura/1286866886.html</guid>
<pubDate>Tue, 12 Oct 2010 12:06:06 +0200</pubDate>
</item>

<item>
<title><![CDATA[Alexandre, el hombre que siempre estuvo allí]]></title>
<link>http://www.elmundo.es/elmundo/2010/10/12/cultura/1286868532.html</link>
<description><![CDATA[&quot;Yo soy actor por Fernando Fernán Gómez. Lo decidí cuando vi a mi amigo recitar
<dc:creator><![CDATA[Mateo Sancho Cardiel (Efe)]]></dc:creator>
<media:description type="html"><![CDATA[&quot;Yo soy actor por Fernando Fernán Gómez. Lo decidí cuando vi a
<media:title type="html"><![CDATA[Uno de los grandes rostros del cine]]></media:title>

<media:content url="http://estaticos.elmundo.es/elmundo/imagenes/2010/10/12/cultura/1286868532_0.jpg" mediu
<media:thumbnail url="http://estaticos.elmundo.es/movil/elmundo/imagenes/2010/10/12/cultura/1286868532_0_10
<guid isPermaLink="true">http://www.elmundo.es/elmundo/2010/10/12/cultura/1286868532.html</guid>
<pubDate>Tue, 12 Oct 2010 09:50:07 +0200</pubDate>
</item>

<item>
<title><![CDATA[La bandera de Venezuela, ausente en el desfile de la Fiesta Nacional]]></title>
<link>http://www.elmundo.es/elmundo/2010/10/11/espana/1286808051.html</link>
<description><![CDATA[La embajada venezolana en Madrid comunicó al Ministerio de Defensa que su abanderado
<dc:creator><![CDATA[Roberto Benito | Agencias]]></dc:creator>
<media:description type="html"><![CDATA[La embajada venezolana en Madrid comunicó al Ministerio de Defensa
<media:title type="html"><![CDATA[Abucheos y gritos de nuevo contra Zapatero]]></media:title>
```

Figura 18 — Información en cada noticia del RSS

Una vez que se han suministrado las direcciones web que la araña recorrerá y el tipo de datos que debe extraer, el siguiente paso es definir cómo hacerlo, y para ello se emplea una biblioteca PHP existente en Internet llamada SimplePie, herramienta que permite leer rápidamente y procesar feeds del tipo RSS, de la que se habló en el apartado anterior.

Para obtener la información que deseamos con SimplePie solo hemos de seguir unos sencillos pasos (**Figura 19**):

- Descargar el archivo con las fuentes primarias [65].
- Copiar el archivo **simplepie.inc** del archivo comprimido que se ha descargado a la carpeta donde se realiza el sistema.

- Se crea un archivo llamado *multifeed.php* donde se definirá el *parser* de feeds a desarrollar.
- Se adjunta SimplePie a la página PHP con la función ***require*.** (1)
- Se crea un objeto del tipo SimplePie que se utilizará para leer el feed. (2)
- Se definirá la ruta del feed que se va a leer, en este caso se le pasa un archivo que contiene todas las URLs a indexar: (3)
    - http://www.elpais.com/rss/feed.html?feedId=1022
    - http://estaticos.elmundo.es/elmundo/rss/portada.xml
    - http://www.larazon.es/rss
    - http://www.abc.es/rss/feeds/abcPortada.xml
    - http://20minutos.feedsportal.com/c/32489/f/478284/index.rs
    - http://www.adn.es/rss/
    - http://www.expansion.com/rss/portada.xml
    - http://www.publico.es/rss/
    - http://estaticos.marca.com/rss/portada.xml
    - http://www.antena3.com/rss/4.xml
    - http://www.europapress.es/rss/rss.aspx?ch=94
    - http://news.google.es/news?pz=1&cf=all&ned=es&hl=es&output=rss

Por lo general, todos los periódicos de ámbito nacional suelen contener las mismas noticias diarias en portada, es por ello, que con este conjunto de URLs se pretende abarcar toda la información representativa en los medios, sin necesidad de indexar más información.

- Se llama a la función ***init***() encargada de iniciar la lectura del feed. (4)
- Adicionalmente se llama a una función ***handle_content_type***() para tener compatibilidad de tipos de codificación de caracteres. (5)

- Y el último paso es obtener la información del feed, (6) por cada una de las noticias publicadas: $title, $content, $url y $fecha, serán las variables con los datos a almacenar en la base de datos.

```php
<?php

(1) require("simplepie.inc");

(2) $feed = new SimplePie();
(3) $feed->set_feed_url(file('multifeed.lst'));

    // Initialize the feed.
(4) $feed->init();

    // Make sure the page is being served with the UTF-8 headers.
(5) $feed->handle_content_type();


    //Store the time 2 hours before
    $date= date('dmygi',strtotime("-2 hour"));


    // Let's loop through each item in the feed.
(6) foreach($feed->get_items() as $item){

    // Let's give ourselves a reference to the parent $feed object for this particular item.
    $feed = $item->get_feed();
    $title = utf8_decode($item->get_title());
    $content = utf8_decode($item->get_content());
    $url = $item->get_permalink();
    $fecha =  $item->get_date('dmygi');
    $fecha2 = $item->get_date('y-m-d g:i');
```

Figura 19 — Código de la araña

Una vez creado el programa que se encarga de extraer la información, se deberá ejecutar automáticamente durante un periodo de tiempo determinado para así recoger la información diaria que se cree en los medios, para ello emplearemos **cron** que es un planificador regular de procesos en segundo plano (demonio) que ejecuta procesos o scripts a intervalos regulares (por ejemplo, cada minuto, día, semana o mes).

El formato de configuración de cron es muy sencillo. El momento de ejecución se especifica de acuerdo con la siguiente tabla (**Figura 20**):

```
#############################################################
#minuto (0-59),                                             #
#|  hora (0-23),                                           #
#|  |  día del mes (1-31),                                 #
#|  |  |  mes (1-12),                                      #
#|  |  |  |  día de la semana (0-6 donde 0=Domingo)        #
#|  |  |  |  |       comandos                              #
#############################################################
15 02  *  *  *
```

Figura 20 — Ejemplo de planificación Cron

4.2.2 Base de datos

Ahora se deberá almacenar toda la información en una base de datos, para su posterior presentación.

Como ya se comentó, en el capítulo 3, se ha utilizado un sistema de bases de datos relacional, hay que recordar que entre las posibilidades barajadas estaba la elección de MySQL frente a PostgreSQL, pero finalmente se eligió MySQL, porque es ideal por su rapidez para bases de datos de pequeño y mediano tamaño, y porque junto con PHP, lenguaje que posee una API para poder conectar con servidores de bases de datos como MySQL, se pueden crear aplicaciones sencillas en poco tiempo.

Una vez tengamos creada la base de datos, podremos acceder y lanzar consultas SQL contra ella mediante PHP (**Figura 21**).

```
//Access to the database

$MAQUINAMYSQL = "localhost";
$USUARIOMYSQL = "dlopez";
$CLAVEMYSQL = "d[illegible]";
$BASEDEDATOS = [illegible];

mysql_connect($MAQUINAMYSQL, $USUARIOMYSQL, $CLAVEMYSQL);
mysql_select_db($BASEDEDATOS);
mysql_query("INSERT INTO Noticias (Titulo,Descripcion,URL,Fecha) VALUES ('$title','$content','$url','$fecha2')");
```

Figura 21 — Código Acceso a BD

Sin necesidad de entrar en detalle de todos los métodos posibles que posee el API de PHP para trabajar con bases de datos, se describen a continuación algunos de los más necesarios:

- **mysql_connect (hostname , username , password)**: establece una conexión a un servidor de MySQL. Todos los argumentos son opcionales, y si no se especifican, los valores por defecto son (' el *localhost'*, nombre del usuario del usuario que posee el proceso del servidor, la contraseña vacía).

- **mysql_select_db()**: establece la base de datos activa que estará asociada con el identificador de enlace especificado. Si no se especifica un identificador de enlace, se asume el último enlace abierto.
- **mysql_query():** envía una sentencia SQL a la base de datos activa en el servidor asociado. Devuelve TRUE (no cero) o FALSE para indicar si la sentencia se ha ejecutado correctamente o no. Un valor TRUE significa que la sentencia era correcta y pudo ser ejecutada en el servidor pero no indica nada sobre el número de filas devueltas. Es posible que la sentencia se ejecute correctamente pero que no devuelve ninguna fila.

Es importante conocer que SQL cuenta con un lenguaje de manipulación de datos (DML) pensado para la consulta, actualización, borrado o inclusión de datos, y que serán las sentencias que debamos usar. La sentencia utilizada para insertar los datos en la tabla será:

```
INSERT INTO Noticias (atributos) VALUES (valores)
```

De esta manera se podrá incluir toda la información necesaria en la base de datos de forma ordenada con el propósito de posteriormente recuperarla en base a unos criterios.

La parte difícil es decidir cómo debería ser la estructura de la base de datos: qué tablas necesitará, y qué columnas habrá en cada tabla. Dado que la aplicación no necesita almacenar información de diferentes clases, con la creación de una sola tabla será suficiente.

La Tabla NOTICIAS, constará de 5 atributos: título, descripción, URL y fecha, que corresponden con la información extraída del procesamiento de datos de los RSS analizados, e Id, que es un campo de tipo autoincremento que se deja a NULL y la base de datos se encargará de generar y asignar los valores consecutivos, de manera que cada noticia tenga un identificador único dentro de la tabla.

Campo	Tipo	Descripción
ID	INT UNSIGNED not null AUTO_INCREMENT	Código identificativo de la noticia
TÍTULO	VARCHAR not null	Título de la noticia
DESCRIPCIÓN	VARCHAR not null	Resumen de la noticia
URL	VARCHAR not null	Link a la página de la publicación
FECHA	VARCHAR not null	Fecha de publicación

Tabla 1 — Relación: Noticias

Como se observa los campos título, descripción, URL y fecha serán de tipo texto, y el identificador de noticia de tipo numérico.

4.2.3 Indexador

Frente a un sistema de recuperación basado en texto como es Lucene o Xapian, capaces de recuperar información de manera automática seleccionando aquellos textos o documentos que son adecuados a una necesidad del usuario, existen sistemas de bases de datos relacionales como MySQL, elegido para el desarrollo del proyecto, que traen incorporado la capacidad de reconocimiento de patrones.

Como ya se ha comentado, MySQL cuenta con un lenguaje de manipulación de datos (DML) con sentencias para consultar, insertar, actualizar y borrar instancias en la base de datos y que serán usadas como sistema de indexación de la información relevante de los RSS analizados.

En primer lugar para almacenar los datos se utiliza la sentencia INSERT, anteriormente descrita, y que inserta nuevos registros en una tabla existente.

Una vez se tiene la información de los RSS almacenada en la base de datos creada, se deberá ser capaz de seleccionar la información que el usuario quiera en cada momento, y para ello se utiliza la sentencia:

```
SELECT * FROM Noticias WHERE (Criterio)
```

```
$b = addslashes(str_replace(' ','%',$buscar));
$criterio = "Fecha BETWEEN CAST('$fechaT' as DATETIME) and CAST('$fechaT2' as DATETIME) AND
(Descripcion LIKE '%".$b."%' OR Titulo LIKE '%".$b."%') ORDER BY Fecha DESC";
$cadbusca=mysql_query("SELECT * FROM Noticias WHERE $criterio");
```

Figura 22 —Uso de la sentencia SELECT

En el código de la aplicación (**Figura 22**) se puede observar exactamente cuál es el criterio usado como selección de la tabla creada. Para la consulta se usa el operador de comparación LIKE seguido de la cadena a buscar entre '%', de esta manera se devolverán todos los artículos que en su titulo o en su contenido aparezca la frase de búsqueda tal y como el usuario la introduzca.

Para permitir que la cadena a buscar contenga más de una palabra, los espacios introducidos en el campo de búsqueda que el usuario rellena en la interfaz de recuperación, se sustituyen también por '%' mediante una función de PHP **str_replace**(), consiguiendo de esta manera que cada palabra sea una búsqueda independiente dentro del texto.

4.2.4 Interfaz de recuperación web

La interfaz de recuperación, encargada del intercambio de información entre el usuario y la base de datos, se ha desarrollado en HTML, lenguaje de marcado para el diseño y estructuración de textos para su posterior representación en forma de hipertexto.

Sin embargo, la evolución de la recomendación HTML sigue la línea de separar el contenido de la presentación, por lo que para representar la información del diseño se ha empleado una hoja de estilo CSS.

Con su utilización se pretende separar el contenido de los documentos de su presentación.

El CSS diseñado funciona a base de reglas, es decir, declaraciones sobre el estilo de uno o más elementos, declarados en la página HTML creada. La regla (**Figura 23**) está formada por dos partes: un selector y la declaración. A su vez la declaración está compuesta por una propiedad y el valor que se le asigne. Así tenemos:

```
body {
    color: white;

    background-color: #A0A0A0;

    font: 15px/21px Arial, Helvetica, sans-serif;
  }
```

Figura 23 — Ejemplo de regla CSS

Donde body es el selector y {color: white; background-color:#A0A0A0; font: 15px} son las declaraciones.

4.2.4.1 Vista principal

Siguiendo las recomendaciones descritas en apartados anteriores se crea la interfaz del buscador: http://plato.it.uc3m.es/~dlopez/proyecto/buscador.php (**Figura 24**).

Figura 24 — Vista Principal: Noticias Diarias

Constará de tres partes bien definidas que serán los campos a rellenar por el usuario para concretar la búsqueda y crear el análisis posterior:

1. Espacio de la búsqueda: donde concretar el tema deseado a buscar.
2. Fecha de inicio: el usuario podrá seleccionar una fecha de inicio a través de un calendario (**Figura 25**), y así decidir a partir de qué día de publicación comenzar a ver resultados de la búsqueda. Por defecto estará seleccionada la fecha de 01/01/2009.
3. Fecha de fin: igual que en el caso anterior se elegirá una fecha de fin de modo que el usuario pueda enmarcar temporalmente el conjunto de noticias mostradas. Por defecto estará seleccionada la fecha correspondiente al día que se esté usando la aplicación.

Figura 25 — Calendario

Una vez rellenados los campos el usuario podrá comenzar la búsqueda tanto pulsando el botón llamado “Buscar”, como pulsando “enter”.

4.2.4.2 Resultados de la búsqueda

Tras introducir los parámetros correctos, se mostrarán todos los resultados que coincidan con los términos seleccionados. El hecho de poder seleccionar una fecha de inicio y otra de fin, permite al usuario filtrar el número de noticias a ser mostradas, y de esa manera solo analizar la información que considere relevante en cada momento.

La página mostrada (**Figura 26**) contendrá tres secciones bien definidas:

- El primer bloque seguirá manteniendo el buscador y la selección de fechas para que el usuario pueda realizar nuevas búsquedas sin tener que volver al menú principal.
- La segunda sección mostrará el sistema de análisis de noticias que se explicará su funcionamiento en siguientes apartados.
- Y finalmente la sección con la información recuperada de la base de datos. Cabe recordar que todos los resultados mostrados contendrán enlaces a la fuente de la noticia original para así, permitir completar al usuario la información mostrada por sistema diseñado.

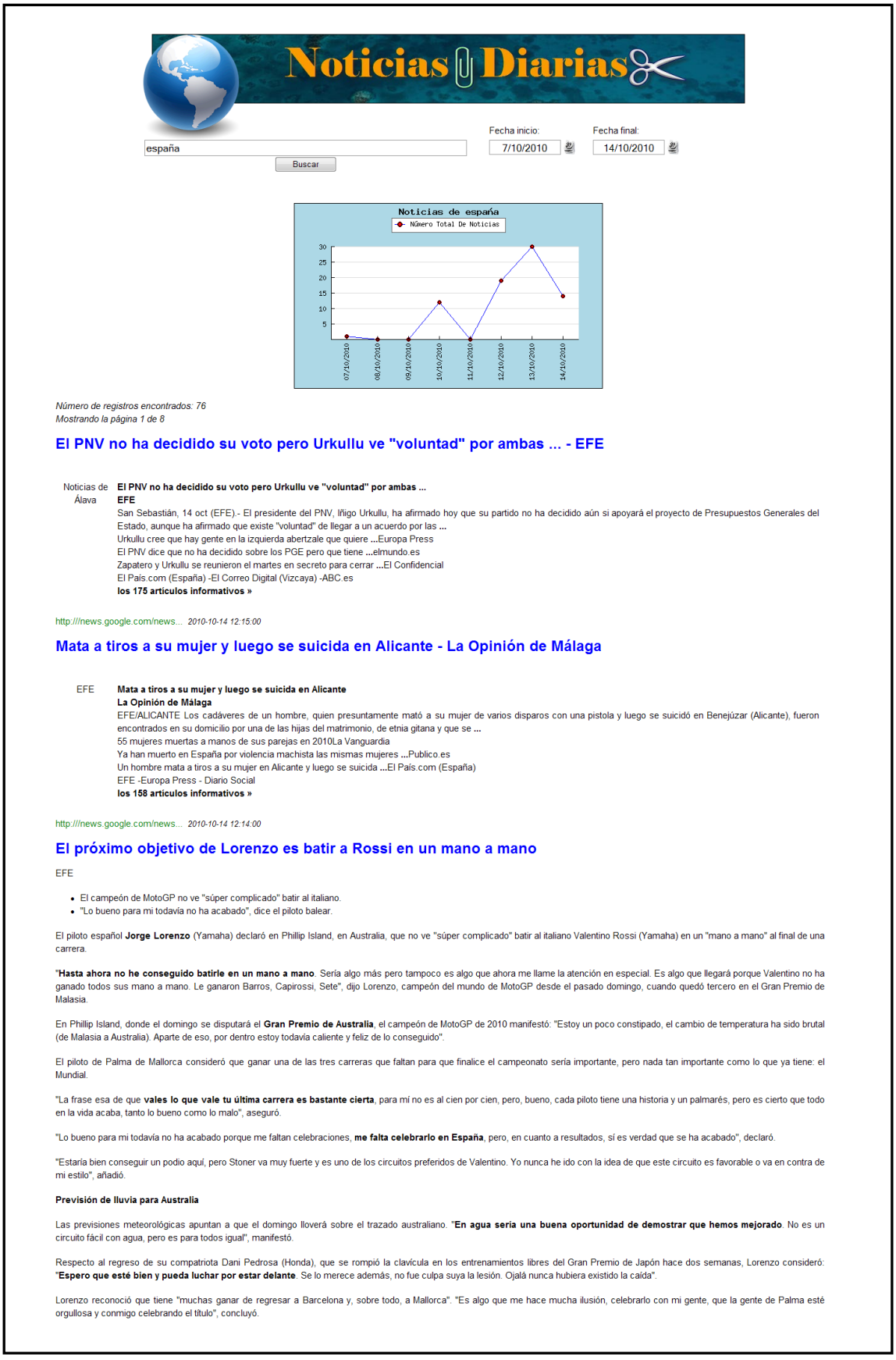


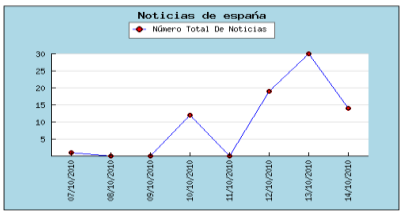
Figura 26.a — Resultado de la búsqueda para el término: España (07/10/10 - 14/10/10)

Noticias Diarias

Fecha inicio: 7/10/2010 Fecha final: 14/10/2010

españa

Buscar



Número de registros encontrados: 76
Mostrando la página 2 de 8

Isabel Pantoja recibe la citación judicial en medio de un circo ... - El Periódico de Catalunya

La Razón **Isabel Pantoja recibe la citación judicial en medio de un circo ...**
El Periódico de Catalunya
Pocas veces se ha visto a la puerta de unos tribunales tal espectáculo. Una auténtica avalancha de periodistas, vecinos indignados y seguidores entusiasmados han acudido a primera hora de la mañana a los juzgados de Marbella para comprobar, in situ, ...
Isabel Pantoja, escoltada por policías, recoge el auto de apertura ...EFE
Isabel Pantoja, con el vestido roto y por la puerta de atrásTerra España
Pantoja necesita escolta policial para entrar en el juzgado por el ...El País.com (España)
20 minutos -Libertad Balear -Diario Directo
los 432 artículos informativos »

http:///news.google.com/news... *2010-10-14 11:25:00*

Los inmigrantes "siguen copando" los trabajos que no quieren los españoles

EFE

- El 40% de los que trabajan en España lo hacen en agricultura o pesca.
- En ambos sectores, representan casi la mitad de los contratados.
- Según un informe de la empresa de trabajo temporal Randstad.

Cuatro de cada diez inmigrantes que trabajan en España lo hacen en la agricultura o en la pesca, según un informe que pone de manifiesto, además, que "siguen copando **los puestos que desestiman los españoles**".

Su presencia en ambos sectores es de tal importancia numérica que representan prácticamente la mitad de los contratados, sobre todo en puestos como peones y lejos de los que exigen una mayor cualificación. El informe, elaborado por la empresa de trabajo temporal Randstad, advierte de que "a pesar de que la crisis y el aumento del paro han provocado una reducción en la oferta laboral, todavía son muchos los trabajadores (españoles) que **seleccionan las oportunidades laborales** que más les interesan".

"Con menos oportunidades y opciones a la hora de acceder a un empleo", los inmigrantes "continúan adscritos -insiste el informe- a aquellos sectores y puestos que no eligen los españoles". Una situación que no es nueva pues incluso en los años de crecimiento económico, **el colectivo de extranjeros** "estuvo relegado a puestos de menor cualificación", al amparo de aquellos sectores con una mayor oferta de mano de obra, como fueron la construcción y la industria.

La inserción laboral de los inmigrantes se da en sectores **"muy específicos"**, sobre todo en la agricultura, la pesca, la minería, la industria manufacturera, la hostelería y el comercio, en puestos "donde apenas es necesaria la formación". Además del 38% que trabaja en la agricultura o la pesca, un 15% lo hace en la industria manufacturera, un 13,75% en la hostelería y un 11,55% en trabajos de limpieza.

El informe de Randstad recalca que mientras **un 21,72% de españoles se emplea en puestos cualificados**, sólo lo consigue un 5,06% de extranjeros. La diferencia es todavía más amplia si se comparan los porcentajes de españoles y extranjeros que ocupan puestos de responsabilidad o de dirección en las administraciones públicas, las empresas o que trabajan como profesionales y técnicos. Un 23,01% frente a un 4,01%.

Los españoles, destaca el informe, "siguen ocupando mayoritariamente puestos de alta cualificación" y sólo aproximadamente uno de cada tres contratados tienen empleos de **baja cualificación**, como trabajos de limpieza, de peón en la agricultura, la pesca, o en la construcción.

"A pesar de la introducción masiva del trabajador inmigrante en el mercado laboral español, todavía existe una gran diferencia entre los puestos que acogen a unos y a otros", insiste el informe, que advierte también de que "los españoles han comenzado a trabajar en puestos de menor cualificación, **reduciendo aún más las posibilidades laborales** de los extranjeros".

http:///20minutos.feedsportal.com/c... *2010-10-14 11:14:00*

El Congreso aprueba que sólo se ejecuten hipotecas tras seis meses de impagos

EFE

- La proposición no de Ley fue presentada por la diputada Rosa Díez y fue consensuada con PP y PSOE.
- La morosidad del crédito hipotecario supera actualmente el 5%, un punto más que el año anterior.
- Esta limitación no sólo se refiere a las hipotecas de los consumidores, sino también a los contratados por empresas y autónomos.

El **Congreso** instó este miércoles al Gobierno a que establezca un mecanismo que limite los **intereses de demora** por deudas **hipotecarias** que deben pagar las familias, con el fin de paliar la **elevada morosidad** que padecen éstas debido a la crisis económica.

De esta forma, la **Comisión de Economía** aprobó por unanimidad la proposición no de Ley presentada por la diputada de UPyD, Rosa Díez, y que ha sido consensuada con el PSOE y el PP a través de dos enmiendas pactadas.

Todos los grupos parlamentarios presentes en la Comisión **apoyaron esta proposición** que pide que no se pueda declarar vencida una operación hipotecaria por el incumplimiento de una sola cuota y sólo pueda ser así cuando se haya producido el impago de **seis cuotas** como mínimo.

La diputada Rosa Díez dijo que se trata de favorecer la economía de las familias y de los consumidores ante unos tipos de interés hipotecarios de bancos y cajas que son "usureros", y recordó que la morosidad del **crédito hipotecario** supera actualmente el **5%** y se ha incrementado en más de un punto este año.

Además, alegó que en las **relaciones** comerciales entre bancos y cajas con las empresas se ha creado un precedente puesto que el **Banco de España** "sí ha establecido limitaciones" sobre la limitación de intereses de demora. "Que no se regule en el caso de los consumidores, que están en peores circunstancias, hace que esta iniciativa sea necesaria y urgente", apostilló.

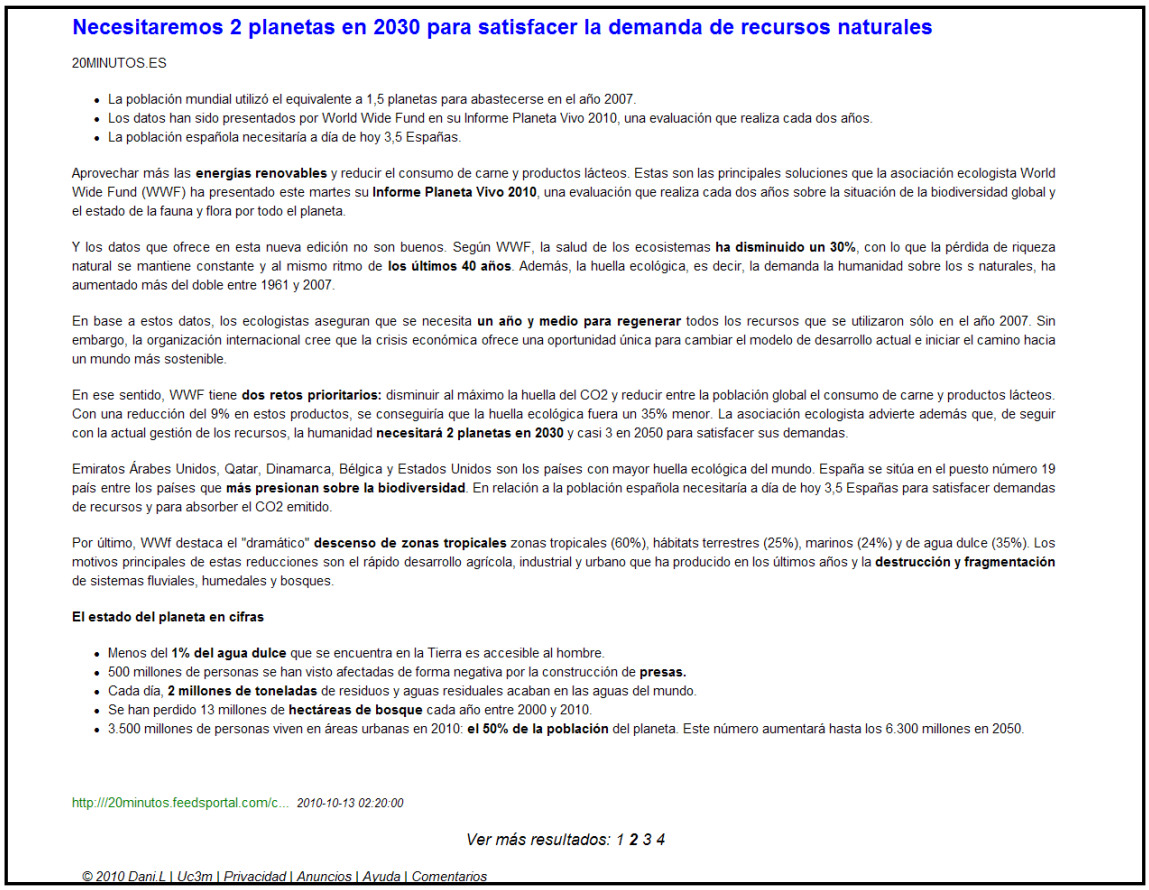
La proposición aprobada urge al Gobierno a que los tipos de interés de demora de prestamos y créditos hipotecarios tengan un límite máximo, situado entre el tipo de interés por descubierto bancario y los intereses de demora comercial. Asimismo, insta a que esta limitación no sólo se refiera a los préstamos hipotecarios de los consumidores sino también a los **contratados por empresas**, **profesionales** o **autónomos**.

Vigilar las prácticas abusivas

La iniciativa, presentada en 2008, también pide que el Gobierno establezca un mecanismo para vigilar las **"prácticas abusivas"** de las entidades financieras.

No obstante, la proposición pactada por los dos grupos parlamentarios mayoritarios de la cámara baja suprime uno de los puntos que pedía UPyD sobre la creación de una nueva estructura administrativa.

Figura 26.b — Resultado de la búsqueda para el término: España (07/10/10 - 14/10/10)



Necesitaremos 2 planetas en 2030 para satisfacer la demanda de recursos naturales

20MINUTOS.ES

- La población mundial utilizó el equivalente a 1,5 planetas para abastecerse en el año 2007.
- Los datos han sido presentados por World Wide Fund en su Informe Planeta Vivo 2010, una evaluación que realiza cada dos años.
- La población española necesitaría a día de hoy 3,5 Españas.

Aprovechar más las **energías renovables** y reducir el consumo de carne y productos lácteos. Estas son las principales soluciones que la asociación ecologista World Wide Fund (WWF) ha presentado este martes su **Informe Planeta Vivo 2010**, una evaluación que realiza cada dos años sobre la situación de la biodiversidad global y el estado de la fauna y flora por todo el planeta.

Y los datos que ofrece en esta nueva edición no son buenos. Según WWF, la salud de los ecosistemas **ha disminuido un 30%**, con lo que la pérdida de riqueza natural se mantiene constante y al mismo ritmo de **los últimos 40 años**. Además, la huella ecológica, es decir, la demanda la humanidad sobre los s naturales, ha aumentado más del doble entre 1961 y 2007.

En base a estos datos, los ecologistas aseguran que se necesita **un año y medio para regenerar** todos los recursos que se utilizaron sólo en el año 2007. Sin embargo, la organización internacional cree que la crisis económica ofrece una oportunidad única para cambiar el modelo de desarrollo actual e iniciar el camino hacia un mundo más sostenible.

En ese sentido, WWF tiene **dos retos prioritarios:** disminuir al máximo la huella del CO2 y reducir entre la población global el consumo de carne y productos lácteos. Con una reducción del 9% en estos productos, se conseguiría que la huella ecológica fuera un 35% menor. La asociación ecologista advierte además que, de seguir con la actual gestión de los recursos, la humanidad **necesitará 2 planetas en 2030** y casi 3 en 2050 para satisfacer sus demandas.

Emiratos Árabes Unidos, Qatar, Dinamarca, Bélgica y Estados Unidos son los países con mayor huella ecológica del mundo. España se sitúa en el puesto número 19 país entre los países que **más presionan sobre la biodiversidad**. En relación a la población española necesitaría a día de hoy 3,5 Españas para satisfacer demandas de recursos y para absorber el CO2 emitido.

Por último, WWf destaca el "dramático" **descenso de zonas tropicales** zonas tropicales (60%), hábitats terrestres (25%), marinos (24%) y de agua dulce (35%). Los motivos principales de estas reducciones son el rápido desarrollo agrícola, industrial y urbano que ha producido en los últimos años y la **destrucción y fragmentación** de sistemas fluviales, humedales y bosques.

El estado del planeta en cifras

- Menos del **1% del agua dulce** que se encuentra en la Tierra es accesible al hombre.
- 500 millones de personas se han visto afectadas de forma negativa por la construcción de **presas.**
- Cada día, **2 millones de toneladas** de residuos y aguas residuales acaban en las aguas del mundo.
- Se han perdido 13 millones de **hectáreas de bosque** cada año entre 2000 y 2010.
- 3.500 millones de personas viven en áreas urbanas en 2010: **el 50% de la población** del planeta. Este número aumentará hasta los 6.300 millones en 2050.

http:///20minutos.feedsportal.com/c... *2010-10-13 02:20:00*

*Ver más resultados: 1 **2** 3 4*

© 2010 Dani L | Uc3m | Privacidad | Anuncios | Ayuda | Comentarios

Figura 26.c — Resultado de la búsqueda para el término: España (07/10/10 - 14/10/10)

Los datos que se muestran en la aplicación son:

- Un resumen del total de noticias encontradas para la búsqueda seleccionada.
- Título de la noticia, que será un enlace a la noticia original mostrada en su propia página web.
- Resumen con imágenes (en el caso de que la noticia tenga).
- Fecha de la publicación.

Todos los resultados son paginados y mostrados en bloques de 20 noticias cada página, además se ordenarán de forma descendente con respecto a su aparición en los medios, es decir, la primera noticia mostrada será la más reciente que se haya publicado, mientras que la última noticia será la más antigua.

4.2.5 Función de análisis de noticias

Finalmente el último módulo que conforma la aplicación web desarrollada es la función de análisis de noticias.

Consistirá en un gráfico (**Figura 27**), que mostrará la evolución en el tiempo de las noticias publicadas en base al término de búsqueda y las fechas de inicio y fin seleccionadas.

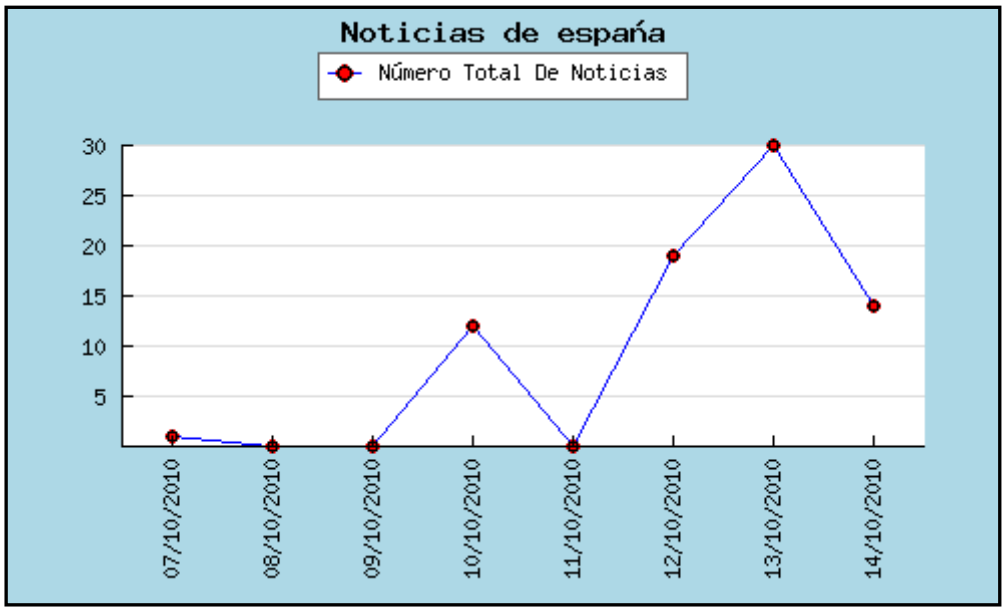


Figura 27 — Gráfico del resultado de la búsqueda del término: España (07/10/10 - 14/10/10)

El **eje x** es el eje temporal, donde cada fecha representa el día en el que se publicó la noticia. Cabe destacar que no todos los días tiene que haber alguna publicación sobre el término buscado, es por ello que el número de resultados devueltos será cero para poder seguir así una distribución homogénea de los datos.

El **eje y** muestra el número total de artículos que se almacenaron relacionados con dicha búsqueda. Así se tiene que el día 07/10/2010 existen 2 noticias relacionadas con el término de búsqueda *España* y sin embargo el día 13/10/2010 hubo 30 resultados para el mismo término.

Con esta distribución se permite al usuario hacerse una idea la relevancia que los artículos mostrados tienen a lo largo de un periodo de tiempo.

Para poder implementar esta gráfica se utiliza una biblioteca de PHP llamada JpGraph que permite crear gráficos matemáticos y estadísticos de manera sencilla y con absoluto control. Además esta biblioteca soporta una amplia variedad de tipos de gráficos para todas las necesidades.

Tras la descarga de los archivos fuente de esta librería en JpGraph Download [75] se crea un archivo PHP llamado grafico.php en cual se incluye (**Figura 28**) primero el archivo jpgraph.php que contiene las funciones básicas, el archivo jpgraph_line.php que contiene las funciones especificas para la creación de gráficos de línea y finalmente el archivo jpgraph_canvas.php que añade funciones para mejorar el diseño.

```
include ("/home/dlopez/public_html/proyecto/src/jpgraph.php");
include ("/home/dlopez/public_html/proyecto/src/jpgraph_line.php");
include ("/home/dlopez/public_html/proyecto/src/jpgraph_canvas.php");
```

Figura 28 — Archivos incluidos en el fichero grafico,php

A continuación el programa se conectará con la base datos para obtener los resultados que concuerden con los parámetros introducidos por el usuario y poder contar el número de resultados devuelto por día.

Y finalmente se pintan los datos siguiendo los ejemplos que posee la biblioteca, pudiendo aplicar los colores y estilos que se deseen. Se puede observar en el código desarrollado (**Figura 29**) que primero se crea una imagen de 500x300 píxeles con los ejes autoescalados, se le dan valores al eje x pasándole como parámetro un array ($x_eje) que contiene el rango de fechas a mostrar y luego se crea un objeto que contiene el gráfico de líneas al que se le pasa como parámetro un array ($ydata) que contiene el número de total de noticias publicadas por día. Posteriormente se agrega este gráfico a la imagen principal y se envía al cliente mediante la función stroke().

```
function yScaleCallback($aVal) {
    return number_format($aVal);
}

$graph = new Graph(500,300,"auto");

$graph->img->SetMargin(60,40,70,40);
$graph->SetScale("textint");
$graph->SetFrame(true);
$graph->SetColor('white');


$graph->title->Set("Noticias de $buscar");
$graph->SetMarginColor('lightblue');

$graph->xaxis->SetLabelAngle(90);
$graph->xaxis->SetTickLabels($x_eje);
$graph->yaxis->SetLabelFormatCallback('yScaleCallback');
$graph->yaxis->SetTextLabelInterval(1);
$graph->yaxis->HideZeroLabel();
$graph->yaxis->SetTitleMargin(55);


$graph->legend->Pos(0.5, 0.08, "center", "top");
$graph->legend->SetLayout(LEGEND_HOR);
$graph->legend->SetShadow(false);
$graph->legend->SetLineWeight(1);
$graph->legend->SetColor('black','darkgray');
$graph->legend->SetFillColor('white');


$p1 = new LinePlot($ydata);
$p1->mark->SetType(MARK_FILLEDCIRCLE);
$p1->mark->SetFillColor("red");
$p1->mark->SetWidth(3);
$p1->SetColor("blue");
$p1->SetCenter();
$p1->SetLegend("Número Total De Noticias");
$graph->Add($p1);


$graph->Stroke();
```

Figura 29 — Código para la creación de la función grafico,php

CAPÍTULO 5. VALIDACIÓN DEL SISTEMA

La validación de este sistema ha tenido lugar de forma paralela a su desarrollo, como factor importante en la optimización del diseño, así como tras la finalización del mismo.

De acuerdo con la estructura del sistema, organizado en torno a dos bloques funcionales, indexación y búsqueda, las pruebas de evaluación se han realizado en ambos. Si bien, debido a la naturaleza de la aplicación, las pruebas efectuadas en la búsqueda han sido de mayor importancia.

5.1 PRUEBAS DE INDEXACIÓN

Para verificar el funcionamiento correcto de la indexación, era necesario comprobar la adecuada inserción de información en la base de datos. Es decir, que todos los campos se estuvieran rellenando adecuadamente y a su vez, que los datos recolectados estuvieran bien codificados.

Una vez revisados todos los datos, se pudo comprobar que la información recolectada tenía un formato utf-8 de codificación debido al proceso de parseo de datos con simplepie, que no era compatible con la codificación ISO-8859-1 de los caracteres originales publicados en los diferentes feeds de datos.

Para solucionar ese problema se empleó una función PHP **utf8_decode**() que decodifica los datos, asumidos a ser codificados por UTF-8, a ISO-8859-1.

5.2 PRUEBAS DE BÚSQUEDA

Las pruebas centradas en el proceso de búsqueda se consideraron como las más importantes en la evaluación global del sistema, debido a que la interacción del usuario con la aplicación vendría dado a través del uso del buscador.

El usuario por regla general deberá introducir, como se explicó en el apartado de diseño, tres parámetros en el formulario de búsqueda para poder realizar una consulta: una o varias palabras clave, fecha de inicio y fecha de fin. Recordar que para facilitar la interacción del usuario, los campos de fechas estarán rellenos automáticamente, si bien el usuario podrá modificarlos como desee.

Para comprobar que el sistema funcionaba adecuadamente se probó que el tipo de parámetros introducidos devolvieran resultados lógicos, además de que si las búsquedas realizadas no obtuvieran coincidencias el usuario estuviera informado del motivo.

El resultado de la evaluación se puede calificar de positivo, ya que no se observaron fallos en la recuperación de información dentro de los límites lógicos. En otras palabras, el sistema es capaz de recuperar información correctamente siempre que cuente con los parámetros introducidos y haya una coincidencia en la base de datos. Para una mayor claridad, se incluyen a continuación algunos ejemplos tomados a partir de las pruebas realizadas.

5.2.1 Ejemplo de prueba 1

Descripción:

Se realiza una búsqueda exclusivamente textual, sin introducir una fecha de inicio o de fin, es decir borrando los datos que vienen introducidos por defecto.

Interpretación de los resultados:

La búsqueda no devuelve resultados e informa al usuario con el mensaje de error: “Por favor introduzca una fecha” el motivo.

5.2.2 Ejemplo de prueba 2

Descripción:

De nuevo se ejecuta una consulta textual, pero en esta ocasión se introduce una fecha de inicio mayor que la fecha de fin.

Interpretación de los resultados:

El buscador nuevamente no devuelve resultados e informa al usuario con el mensaje de error: “La fecha inicial debe ser menor que la fecha final”.

5.2.3 Ejemplo de prueba 3

Descripción:

Se realiza una búsqueda sin introducir texto, exclusivamente con los valores de fechas que vienen por defecto en la aplicación.

Interpretación de los resultados:

El sistema no devuelve ninguna búsqueda, e informa al usuario que debe introducir alguna palabra clave para poder comenzar a obtener resultados.

Figura 30 — Detalle de prueba Nº3

5.2.4 Ejemplo de prueba 4

Descripción:

Se lleva a cabo a cabo una consulta combinada, en la que se introduce una palabra clave: “social” y se mantienen las fechas que la aplicación devuelve por defecto.

Interpretación de los resultados:

La aplicación de búsqueda devuelve correctamente resultados que contienen la palabra clave y además que están comprendidos entre las fechas introducidas. Además los resultados están bien ordenados por fecha de publicación y el gráfico muestra la cantidad de noticias registradas cada día.

5.2.5 Ejemplo de prueba 5

Descripción:

Se ejecuta una consulta con más de una palabra clave: “Zapatero, Crisis”, además de introducir manualmente las fechas mediante el uso del calendario asociado.

Interpretación de los resultados:

Nuevamente la búsqueda devuelve correctamente los resultados coincidentes con la información de la base de datos.

5.2.6 Ejemplo de prueba 6

Descripción:

Se ejecuta la misma consulta anterior, pero se sustituye una de las palabras clave por: “Ventana”.

Descripción de los resultados:

En este caso el sistema muestra un mensaje: “No se encontró ninguna búsqueda relacionada”, dado que no hay ninguna coincidencia con la información de la base de datos.

CAPÍTULO 6. PRESUPUESTO

UNIVERSIDAD CARLOS III DE MADRID
Escuela Politécnica Superior

PRESUPUESTO DE PROYECTO

1.- Autor:

Daniel López Fuentes

2.- Departamento:

Ingeniería Telemática

3.- Descripción del Proyecto:

- Titulo: **Sistema de seguimiento y análisis de medios de comunicación en Internet**
- Duración (meses): **12**
- Tasa de costes Indirectos: **20%**

4.- Presupuesto total del Proyecto (valores en Euros):

22.330,00 € Euros

5.- Desglose presupuestario (costes directos)

PERSONAL

Apellidos y nombre	N.I.F.	Categoría	Dedicación (personas mes) [a]	Coste hombre mes	Coste (Euro)
Villena Román, Julio		Ingeniero Senior	0,5	4.289,54	2.144,77
López Fuentes, Daniel		Ingeniero	6	2.694,39	16.166,34
		Personas mes 6,5		**Total**	18.311,11

[a] 1 Persona mes = 131,25 horas. Máximo anual de dedicación de 12 personas mes (1575 horas)
Máximo anual para PDI de la Universidad Carlos III de Madrid de 8,8 personas mes (1.155 horas)

EQUIPOS

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable [d]
Servidor	1.500,00	100	12	60	300,00
Disco de Backup	150,00	100	12	60	30,00
				Total	330,00

[d] Fórmula de cálculo de la Amortización:

$$\frac{A}{B} x C x D$$

A = nº de meses desde la fecha de facturación en que el equipo es utilizado
B = periodo de depreciación (60 meses)
C = coste del equipo (sin IVA)
D = % del uso que se dedica al proyecto (habitualmente 100%)

SUBCONTRATACIÓN DE TAREAS

Descripción	Empresa	Coste imputable
	Total	0,00

OTROS COSTES DIRECTOS DEL PROYECTO[e]

Descripción	Empresa	Costes imputable
	Total	0,00

[e] Este capítulo de gastos incluye todos los gastos no contemplados en los conceptos anteriores, por ejemplo: fungible, viajes y dietas, otros,...

6.- Resumen de costes

Presupuesto Costes Totales	**Presupuesto Costes Totales**
Personal	18.311
Amortización	330
Subcontratación de tareas	0
Costes de funcionamiento	0
Costes Indirectos	3.662
Total	22.303

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

Antes de empezar un proyecto es importante crear y declarar una serie de objetivos que se deben cumplir. Dichos objetivos fueron desde el principio la creación de una herramienta web capaz de hacer un seguimiento y análisis de medios de comunicación que ofreciera al usuario la facilidad de buscar información y a la vez gestionarla, y que durante el apartado de desarrollo del proyecto se trató de dar cuenta de los pasos seguidos para su logro. Corresponde a este apartado someter a evaluación dichos objetivos para así poder concluir su correcta consecución.

Como se ha expuesto a lo largo de la memoria, Internet se ha convertido en una fuente inagotable de información a la que millones de usuarios acceden diariamente, y donde la Web 2.0 supone la constatación de que es esencial la posibilidad de compartir, crear, generar o difundir contenidos, informaciones y datos desde cualquier lugar en cualquier momento.

Un estudio reciente revela que la gestión de la información en las empresas, hace que los empleados utilicen 14,5 horas a la semana a leer y contestar e-mails; 13,3 horas creando documentos y 9,5 horas analizando información, lo que supone que una empresa con mil empleados pierde 5,7 millones de dólares (4,3 millones de euros) anualmente en el tiempo que sus trabajadores necesitan para adaptar la información a las diferentes aplicaciones tecnológicas [74].

A esto, se añade que la gestión de prensa digital se hace cada día una tarea más compleja. Diariamente el número de periódicos y webs aumenta y la prisa por tener la información a primera hora hace que cualquier profesional o empresa que necesite conocer

puntualmente la información que en los medios de comunicación se publica, tenga en la confección de un resumen diario de prensa, uno de sus principales retos.

Es por ello que uno de los requisitos y objetivos cumplidos durante la creación del proyecto sea permitir al usuario buscar fácilmente entre todas las noticias almacenadas en la base de datos creada, y de ese modo encontrar aquellas que son necesarias en cada momento.

Por otra parte, las noticias son efímeras, pierden valor al envejecer. Los lectores quieren enterarse ahora. Lo que pasó esta mañana o ayer tiene más valor noticioso que lo que pasó la semana pasada. Sin embargo, un nuevo ángulo, descubrimiento o revelación puede rejuvenecer una noticia vieja. La misma lógica se aplica a las noticias futuras. La noticia de que habrá una huelga general mañana tiene más valor informativo, que la noticia de que la huelga se producirá la semana próxima.

En conclusión: mientras más cercana esté la noticia a su fecha de publicación, más valor tendrá para el usuario. Por ello generar automáticamente un gráfico que analice la información mostrada, otro de los objetivos del proyecto, proporciona al usuario un análisis de la relevancia que posee esa noticia y aporta una visión más general de su importancia en los medios elegidos.

En último lugar, otro objetivo que se quiso cumplir durante la realización del proyecto, fue tratar de crear una aplicación accesible y fácil de usar. Para ello, se definió la estructura de una página web funcional, optimizando al máximo su arquitectura, de manera que se agilizara la navegación por parte del usuario.

Finalmente, dado que la aplicación no está enfocada a ningún sector en concreto, cualquier empresa o particular que necesite clasificar las noticias que aparecen diariamente en los periódicos publicados en Internet y a la vez hacer un análisis sobre esta información, conociendo la repercusión de una noticia en los medios, podrá usarla.

7.2 TRABAJOS FUTUROS

Existen varias líneas de trabajo que se han ido planteando en la realización de este proyecto y en las que se podría seguir investigando. A continuación se detallan algunas de ellas:

- Resultaría interesante emplear la herramienta de seguimiento para hacer una comparativa de varias noticias a la vez de manera que el usuario pudiera tener una visión más global de cómo la información de actualidad gana o pierde relevancia con el tiempo, algo parecido a lo que hace la aplicación web: News Brief [53].
- Por otra parte dado que se trata de una aplicación "sencilla", cabe destacar que para que el resultado de la búsqueda permita llegar a los datos que se quieren hallar, es fundamental aprender a escribir correctamente la pregunta que se enviará al buscador. Para ello, primero el usuario deberá preguntarse qué palabras clave pueden estar relacionadas con el tema deseado. Esto, por supuesto, no siempre es sencillo pero tampoco imposible. Esto es algo que se podría mejorar con sistemas de indexación más avanzados capaces de clasificar mejor la información, como Lucene o Xapian, descritos en el apartado de tecnologías.
- También sería interesante, dado la cantidad de navegadores web y sus respectivas versiones que existen en la actualidad en el mercado, optimizar la aplicación para que pudiera "verse" bien en cualquiera de ellos, esto es una tarea bastante difícil de lograr, de manera que por el momento la aplicación esta optimizada para los navegadores Explorer, versión 7 y 8, y Firefox versión 3.5 y 3.6. Para el resto de navegadores aunque no cambia mucho el aspecto que muestra la web puede sufrir alguna variación de diseño.

- De manera similar a lo anterior, en la actualidad el mundo de la tecnología móvil crece a pasos agigantados y sus navegadores deben hacer frente a la interfaz que proporcionan dichos dispositivos, es por ello que también sería aconsejable trabajar en el diseño de la página para su correcta visualización en estos dispositivos. Por otro lado además este tipo de dispositivos poseen lo que se conoce como aplicaciones nativas que son propias de cada sistema operativo, así los teléfonos HTC utilizan aplicaciones Android, o los modelos de Apple (Iphone, Ipad) utilizan aplicaciones propias.

  Esto genera un nuevo mercado en el que las aplicaciones web van perdiendo la partida y es por esto que crear una aplicación nativa que realizara la misma funcionalidad descrita en la aplicación desarrollada establecería nuevos mercado de comercialización.
- Otro trabajo futuro sería poder crear una base de datos más amplia donde no sólo se almacenaran noticias de actualidad española sino que se recogiera información de periódicos de todo el mundo, lo cual conllevaría quizás también añadir periódicos que no publicaran sus noticias en feeds RSS, y por consiguiente se debería mejorar la araña web diseñada a fin de poder recolectar toda la información útil.
- Finalmente para completar el sistema, se podrían crear módulos de análisis más avanzados, como agrupamiento de noticias, descubrimiento de "*trending topics*" (temas del momento), análisis de opinión, etc.

BIBLIOGRAFÍA Y REFENCIAS

Las referencias bibliográficas se encuentran ordenadas alfabéticamente por apellidos de autor; las referencias localizadas en URL aparecen en segundo término, ordenadas por orden de aparición durante la lectura de la memoria.

1. Recursos bibliográficos

[1] ACEVEDO Fernando, David ZURDO Y Alejandro SICILIA, "Buscadores de Internet", Paraninfo, 1998.

[2] BAEZA-YATES, R. y B. RIBEIRO-NETO, "Modern Information Retrieval", Addison Wesley, 1999.

[3] BELKIN, N. J., "User interfaces for information systems", Revista española de Doc. Científica, vol.14, nº2, 1991, pag.193-213.

[4] CASTELLS, Pablo, "La web semántica", en C. Bravo y M. A. Redondo (coord.): *Sistemas interactivos y colaborativos en la web,* Cuenca: Universidad de Castilla la Mancha, 2005.

[5] CHEN, H. et al., "Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques", Journal of the American Society for the Information Science, nº49, 7, 1998, pag. 541-556.

[6] COLLADA PÉREZ, Sonia, "Sistema de indexación y búsqueda de documentos audiovisuales", Proyecto Fin de Carrera, 2009.

[7] DUBOIS, Paul, "Edición Especial MySQL", Prentice Hall, 2001.

[8] ELMASRI, Ramez y Shamkant B. NAVATHE, "Fundamentos de Sistemas de Bases de Datos", Pearson, 2007.

[9] FALKNER, Jayson, "Desarrollo Web con JSP: Fundamentos", Anaya, 2002.

[10] FORNAS CARRASCO, Ricardo, "Gobib: una guía de buscadores en Internet", Revista Métodos de Información, v. 4, nº 21, p. 28-31, 1997.

[11] GARCÍA, Francisco Javier, "Desarrollo de un Web Log de encuestas en PHP", Proyecto fin de carrera, 2002.

[12] GARZÓN CASADO, Ruth, "Desarrollo de mejoras a un conversor de HTML a XHTML: Conversión a XHTML basic", Proyecto fin de carrera, 2008.

[13] GILSTER, P., "Finding It on the Internet. The Internet´s Navigator Guide to Search Tools and Techniques", Nueva York, John Wiley, 1996.

[14] GLASS, Michael k., "Desarrollo Web con PHP, Apache y MySQL: Fundamentos", Anaya, 2004.

[15] INGWERSEN, P., "Information Retrieval Interaction", London, Taylor Graham, 1992.

[16] LÓPEZ HERRERA, Antonio Gabriel, "Modelos de Sistemas de Recuperación de Información Documental Basados en Información Lingüística Difusa", Tesis Doctoral, Universidad de Granada, 2005.

[17] MACIÁ DOMENE, Fernando y Javier GOSENDE GRELA, "Posicionamiento en buscadores", Anaya, 2007.

[18] MANRIQUE DE LARA CADIÑANOS, Miguel, "Herramienta para la anotación ontológica de noticias en formato NITF", Proyecto fin de carrera, 2006.

[19] MASLAKOWSKI, Mark, "Aprendiendo MySQL en 21 días", Prentice Hall, 2001.

[20] MUKHAR, Kevin, Todd LAUINGER, y John CARNELL, "Bases de datos con Java: Fundamentos", Anaya, 2002.

[21] OLMEDA GÓMEZ, Carlos, "Apuntes de la asignatura: Técnicas de Búsqueda y Uso de la Información", Universidad Carlos III, Madrid.

[22] ORÓS CABELLO, Juan Carlos, "Diseño de páginas Web con XHTML, JavaScript y CSS", Ra-ma, 2005.

[23] RUIZ GUIJARRO, Francisco Javier, "Desarrollo de un sistema de búsqueda web basada en categorías y foros de noticias", Proyecto fin de carrera, 2005.

[24] SALAZAR, Idoia, "Las profundidades de Internet: accede a la información que los buscadores no encuentran y descubre el futuro inteligente de la Red", Ediciones Trea, 2005.

[25] SALTON, G., A. WONG y C. S. YANG, "A Vector Space Model for Automatic Indexing", Commun. ACM, Vol.18, Nº11, pp 613-620, noviembre 1975.

[26] TRAMULLAS SAZ, Jesús y María Dolores OLVERA LOBO, "Recuperación de la información en Internet", Ra-ma, 2001.

[27] TRAMULLAS SAZ, Jesús, "Introducción a la Documática, 1: Teoría", Zaragoza, Kronos, 1997.

[28] ULLMAN, Larry, "PHP", Prentice Hall, 2001.

[29] VALIENTE BLÁZQUEZ, María, "Desarrollo de un buscador adaptativo", Proyecto de fin de carrera, 2003.

[30] WEBSTER, K. y K. PAUL, "Beyond Surfing: Tools and Techniques for Searching the Web", Information Technology, enero 1996.

2. Recursos en línea

[31] Internet, Medio Favorito para Leer las Noticias,

http://www.contactomagazine.com/internetvsprensa1006.htm

[visitado el 04/03/2010]

[32] José Luis Salmerón Silvera, "Localización de información en motores de búsqueda en Internet. Análisis de la efectividad", Economía Industrial, Nº364, 2002.

http://www.mityc.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/346/15%20SALMERON.pdf

[visitado el 04/03/2010]

[33] Ifra técnicas de prensa, "Web 2.0 edición especial",

http://www.nxtbook.fr/nxtbooks/ifra/web2-0_stdp/index.php

[visitado el 04/03/2010]

[34] CNN, http://edition.cnn.com/ [visitado el 12/04/2010]

[35] Wikipedia, http://es.wikipedia.org/wiki/Wikipedia:Portada

[visitado el 10/03/2010]

[36] Jesús Tramullas Saz, Los sistemas de Información: una reflexión sobre información, sistema y documentación, Revista General de Información y Documentación, Vol.7, nº1, 1997.

http://revistas.ucm.es/byd/11321873/articulos/RGID9797120207A.PDF

[visitado el 12/04/2010]

[37] Ángeles Maldonado Martínez y Elena Fernández Sánchez, "Análisis comparativo de buscadores en internet", El profesional de la información, vol.9, nº3, 2000.

http://www.elprofesionaldelainformacion.com/contenidos/2000/marzo/5.pdf

[visitado el 09/03/2010]

[38] Recuperación de información,

http://ict.udlap.mx/people/carlos/is215/ir12.html [visitado el 15/04/2010]

[39] Recuperación y Organización de la información, Universidad Carlos III de Madrid, http://modelosrecuperacion.tripod.com/ [visitado el 15/05/2010]

[40] Yahoo! España, http://es.yahoo.com/?p=us [visitado el 18/06/2010]

[41] Google España http://www.google.es [visitado el 18/06/2010]

[42] Sitios más visitados de Internet,

http://www.rankeen.com/Rankings/rank_sitios_visitados.php

[visitado el 18/06/2010]

[43] Top 500 sites on the web, http://www.alexa.com/topsites

[visitado el 19/06/2010]

[44] Dmoz (Open Directory Project), http://www.dmoz.org/

[visitado el 13/06/2010]

[45] Galaxy, http://www.galaxy.com/ [visitado el 13/06/2010]

[46] Bing, http://www.bing.com/ [visitado el 18/06/2010]

[47] Comparativa de resultados entre Google y Yahoo! para una misma Keyword, http://langreiter.com/exec/yahoo-vs-google.html [visitado el 10/06/2010]

[48] Google features, http://www.google.es/help/features.html

[visitado el 18/06/2010]

[49] Búsquedas en Google: Hacia la web semántica,

http://www.corbax.com/blog/busquedas-google-web-semantica-web-3-0/

[visitado el 18/06/2010]

[50] Entienda la web 2.0 y sus principales servicios,

http://www.eduteka.org/Web20Intro.php [visitado el 15/05/2010]

[51] Rollyo, http://www.rollyo.com [visitado el 21/06/2010]

[52] Como crear un RSS, http://www.fullpracticos.com/rss_como_crear_3.html

[visitado el 02/07/2010]

[53] News Brief, http://press.jrc.it [visitado el 05/04/2010]

[54] Tus titulares, http://www.tustitulares.com [visitado el 10/07/2010]

[55] Spy Press, http://www.spypress.com [visitado el 10/07/2010]

[56] Googlebot, http://google.dirson.com/googlebot.php

[visitado el 18/06/2010]

[57] W3C: World Wide Web consortium "Leading the web to its full potential…" Consorcio internacional para el desarrollo de tecnologías Web, http://www.w3.org

[visitado el 12/07/2010]

[58] HTML 4.01 Especificación, http://www.w3.org/TR/html401/

[visitado el 12/07/2010]

[59] Extensible Markup Language (XML) 1.0, http://www.w3.org/TR/xml/

[visitado el 12/07/2010]

[60] XHTML 1.0, recomendación del W3C, http://www.w3.org/TR/xhtml1/

[visitado el 12/07/2010]

[61] Cascading Style Sheets, level 1 & 2, Specification, http://www.w3.org/TR/REC-CSS1/ & http://www.w3.org/TR/CCS2/

[visitado el 13/07/2010]

[62] JavaScript, http://www.w3schools.com/js/default.asp

[visitado el 15/07/2010]

[63] CGI: Common Gateway Interface, http://www.w3.org/CGI/

[visitado el 15/07/2010]

[64] PHP, http://www.php.net [visitado el 13/07/2010]

[65] SimplePie, http://simplepie.org/ [visitado el 05/04/2010]

[66] JSP, http://geneura.ugr.es/~jmerelo/JSP/ [visitado el 17/07/2010]

[67] ASP-NET,

http://www.es-asp.net/tutoriales-asp-net/tutorial-61-62/empezando.aspx

[visitado el 17/07/2010]

[68] JDBC, http://www.scribd.com/doc/3321228/JDBC [visitado el 18/07/2010]

[69] PostgreSQL, http://www.postgresql-es.org/ [visitado el 17/07/2010]

[70] Swish-e, http://swish-e.org/ [visitado el 21/07/2010]

[71] Xapian, http://www.xapian.org/ [visitado el 21/07/2010]

[72] Lucene, http://lucene.apache.org/ [visitado el 21/07/2010]

[73] HCI, Interfaz de un buscador,

http://hciinterfazbuscador.iespana.es/interfazbuscadorejemplosdiseno.html

[visitado el 02/08/2010]

[74] Datos de internet del 2006,

http://www.noticiasdot.com/wp2/2007/03/07/los-datos-de-internet-del-2006-

equivalen-a-12-pilas-de-libros-de-la-tierra-al-sol/ [visitado el 25/08/2010]

[75] JpGraph http://jpgraph.net/download/index.php [visitado el 17/08/2010]