Universidad de las Ciencias Informáticas

Facultad 1

Subsistema de búsqueda de noticias del buscador Orión

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

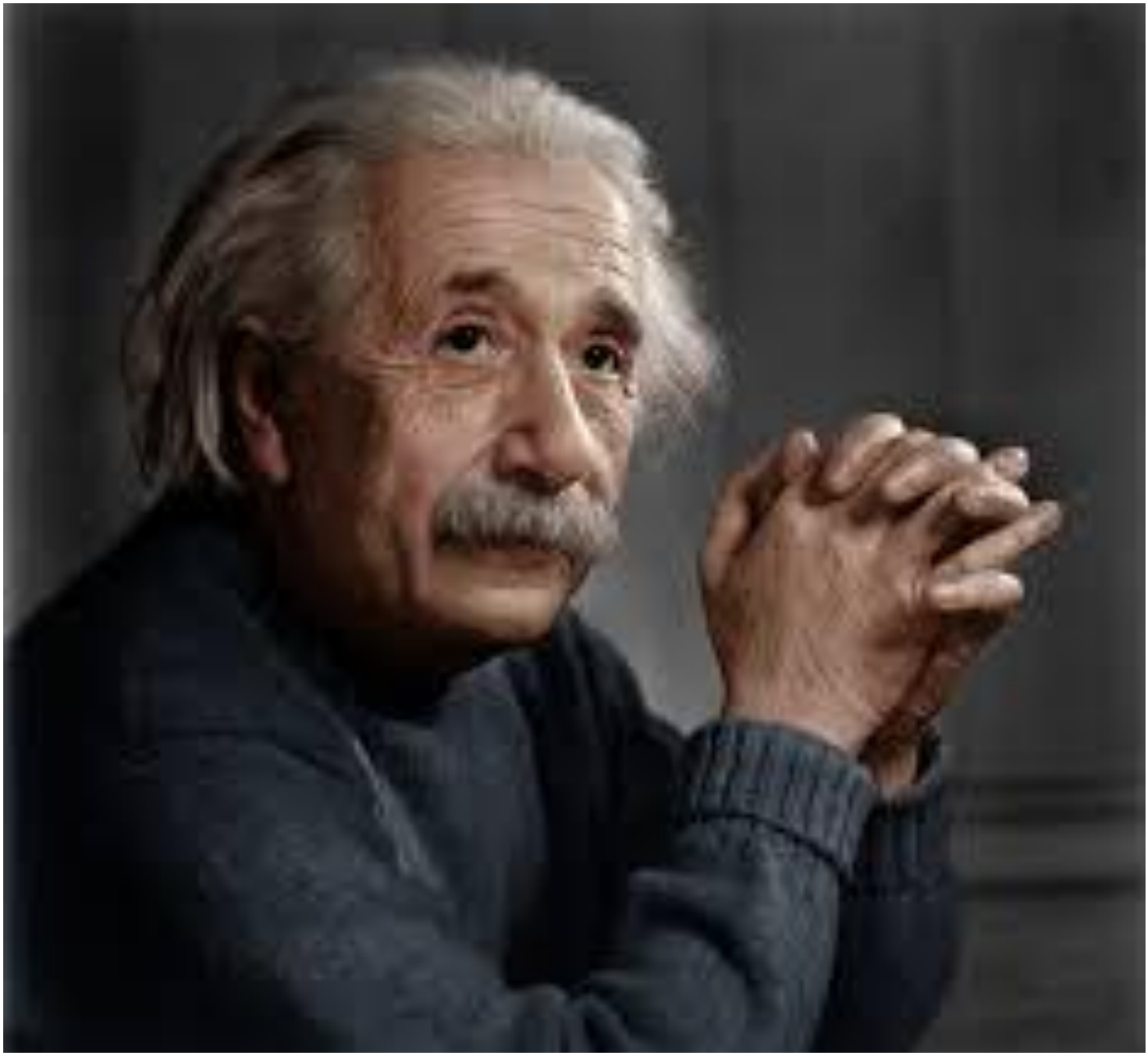
Autor:

Amaury Jorge Reynaldo

Tutor(es):

Ing. Martha Luisa Gala Rodríguez
Ing. Odisleysi Martínez Furones
Ing. Roannel Fernández Hernández

La Habana, Junio 2016



"Hay una fuerza motriz más poderosa que el vapor, la electricidad y la energía atómica: la voluntad"

Albert Einstein

Declaración de autoría

Declaro ser el autor de la presente tesis nombrada Subsistema de búsqueda de noticias del buscador Orión, reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año _____.

Amaury Jorge Reynaldo.

_____

Firma del autor

Ing. Martha Luisa Gala Rodríguez

_____

Firma del tutor

Ing. Odisleysi Martínez Furones

_____

Firma del tutor

Ing. Roannel Fernández Hernández

_____

Firma del tutor

Resumen

Para localizar y procesar la gran cantidad de información existente en Internet de forma rápida y automática son utilizados los motores de búsqueda. Estos sistemas para ser completos, deben contar con opciones que permitan la búsqueda única de información como por ejemplo noticias. Sin embargo, actualmente el buscador cubano Orión no cuenta con una sección que solo muestre noticias digitales. Por tal motivo, la presente investigación propone desarrollar el subsistema de búsqueda de noticias del buscador Orión, con el objetivo de optimizar las búsquedas de noticias digitales publicadas en el dominio .cu. La propuesta de solución está compuesta por tres componentes: Rastreador, Indexador y Aplicación web; los cuales, permiten a los usuarios realizar búsquedas simples y avanzadas, filtrando los contenidos atendiendo a criterios previamente definidos. Para ello se desarrolló el análisis, diseño, implementación y prueba del subsistema que constituye el resultado principal de la investigación desarrollada. Conjuntamente se emplearon varias herramientas y una metodología para guiar el desarrollo de la solución.

Palabras clave: *aplicación web, búsqueda, indexador, noticia, rastreador.*

Contenido

Índice de ilustraciones

Índice de tablas:

Introducción

Con la expansión y desarrollo de la sociedad, la humanidad ha tenido la necesidad de hacer uso de los conocimientos disponibles en el mundo para resolver problemas científicos, sociales o de otra índole. Lo que trajo consigo un nuevo desafío para los hombres: encontrar una forma de acceder a toda la información disponible de una manera rápida y sencilla. A partir de esta necesidad y teniendo en cuenta los avances tecnológicos logrados por el hombre surge Internet en 1982. (Fernández 2008) , primeramente, con objetivos militares, pero dada la magnitud del descubrimiento, pasa a ser de dominio público convirtiéndose en la principal vía de acceso a la información. Posteriormente, surge la Web[1] como parte de Internet, la cual también evoluciona hasta convertirse en la Web 2.0.

La Web 1.0 fue el principio del inicio del desarrollo de las telecomunicaciones, el usuario era restringido a leer la misma información ya que no se podía actualizar continuamente. Con la evolución a la Web 2.0 se tiene un gran avance como es el intercambio de información entre usuarios y la interacción en redes sociales como el Facebook y entre otros, permitiendo la comunicación con todo el mundo.

El concepto de Web 2.0 se desprende del antiguo modelo donde los usuarios eran meros consumidores de información. Convirtiéndose en un espacio donde cambia el rol del usuario de consumidor a creador, ya que esta nueva forma permite navegar e interactuar de manera dinámica con los contenidos existentes, además de permitir crearlos. Los cambios ocurridos en la forma o naturaleza de los usuarios y el manejo de la información son debido a la posibilidad de intercambiar contenidos, opiniones y aportar en la creación de información nueva. (Pupo Gómez et al. 2014). La forma que adopta la Web 2.0 es una manera de hacer que la organización y la dirección de la información dependan de cómo actúan las personas que acceden a ella.(Rey 2013).

Con la Web 2.0, han surgido varios servicios, entre ellos se destacan los servicios de información[2] que propician información a los usuarios. Así como, servicios interactivos donde millones de usuarios en la red comparten sus informaciones. Muestra de esos servicios son los medios de comunicación que han aparecido por todo el Internet: agencias de noticias, periódicos y revistas que en su conjunto comparten el objetivo transmitir mensajes en forma de noticia a los usuarios.

[1] Es un sistema de distribución de documentos de hipertexto o hipermedios interconectados y accesibles vía Internet (Monografias.com 2016)

[2] Actividades identificables e intangibles, que el profesional de la información ofrece al usuario

La evolución de los medios de comunicación ha estado siempre estrechamente vinculado al desarrollo de las tecnologías y en especial, a los avances en el terreno de las telecomunicaciones; han hecho posible no solo que una noticia ocurrida en algún lugar del mundo pueda ser conocida casi instantáneamente, sino que también han asegurado una difusión amplia y rápida de la información hacia el público, además de estar presentes de forma significativa en la propia redacción de las noticias (Solana 1988). Con la metamorfosis de los medios surgen los llamados Medios de comunicación masiva.

Los medios de comunicación masivos, son aquellos en que se envían mensajes por un emisor y se reciben de manera idéntica por varios grupos de receptores, teniendo así una gran audiencia; el mundo los conoce y reconoce como la televisión, la radio o el periódico (Domínguez Goya, 2012). Entre sus principales características se encuentra que son abiertos, lo que significa que cualquier persona puede acceder a ellos, permitiendo influir sobre la sociedad de manera considerable, ayudando a satisfacer sus necesidades ya sea informar, educar, entretener, formar opinión o publicidad y propaganda (“Medios de comunicación masiva” 2015). Por lo cual, una de las formas de acceso a la información más utilizadas actualmente es a través de la noticia.

La noticia es un hecho verdadero, inédito y actual, de interés general, que se comunica a un público que puede considerarse masivo, una vez que ha sido recogido, interpretado y valorado por los emisores que controlan el medio utilizado para la difusión del mensaje (Cuadrado, 2012). Este tipo de información es manejada por los medios de comunicación, en especial los medios de comunicación masivos, entre ellos se encuentra Internet que se ha convertido en un gran medio de comunicación en el mundo.

Se puede afirmar que internet es un gran medio, analizando la cantidad de usuarios que lo usan actualmente, estimado en 3.200 millones según (Quintana 2016), además, de su concepto, el cual plantea que un medio de comunicación es una vía por la cual se transmite un mensaje. (Arias Molina, 2011). Y teniendo en cuenta que Internet es un conjunto de ordenadores y servidores conectados entre sí trasmitiendo mensajes, entonces se cumple la anterior afirmación. Además, cuenta con diferentes expresiones como visuales, escritas, sonoras y audiovisuales. Así también incluye la posibilidad de una comunicación interpersonal, grupal y masiva (Crovi Druetta, 2006).

Aprovechando las posibilidades de Internet y haciendo uso de sus principales características tales como la inmediatez[3] y posibilidad de difusión mundial, la sociedad comenzó a hacer un mayor uso de este medio de

[3] Capacidad de publicación o recuperación de información de manera rápida.

comunicación. Debido a que cada día suceden múltiples acontecimientos de diversos orígenes, ya sean militares, económicos, de salud, deportivos o de entretenimiento, surge la necesidad de difundir estos hechos para mantener a las personas actualizadas sobre el acontecer mundial. Para lograr el objetivo de mantener a los usuarios informados se requiere de una propagación casi instantánea de la información, y como se ha mencionado Internet provee la forma adecuada para hacerlo.

Los hechos o acontecimientos obtienen el nombre de noticia en el siglo XIX cuando algunos diarios dividieron la información (los hechos) de la opinión (los comentarios) es lo que da nacimiento a la noticia como género periodístico, cumpliendo con el requisito indispensable de contestar las preguntas básicas de la información: qué, quién, dónde, cuándo y cómo (Castro, Olivares, y Martínez 2014). La misma desempeña un papel importante en el desarrollo de la sociedad, puesto que es un medio de apoyo a la toma de decisiones en el ámbito gubernamental, empresarial, científico o simplemente en la vida cotidiana de la sociedad. Debido a la importancia de las noticias varios países cuentan con sus propios medios periodísticos u otros medios que publican noticias. Evidencia de esto es *The New York Times* en Estados Unidos, El País en España o la BBC[4] en el Reino Unido.

Cuba no está exenta de este desarrollo, evidencia de ellos es que en la red nacional han surgido numerosos sitios web informativos, como revistas y periódicos. Según Cubadebate[5] se tienen al menos 58 sitios web de tipo noticioso bajo el dominio .cu (21 periódicos nacionales o provinciales, 2 agencias de información y noticias, 13 sitios correspondientes a canales de televisión, 6 revistas y 16 emisoras de radio) donde se encuentran numerosas publicaciones de carácter nacional e internacional. Además, de contar con bitácoras (*blogs*) personales donde también se registran publicaciones y que están publicados en el mismo servidor que el periódico al que están adscritos.

Debido a la cantidad de información contenida en la Web con diversos formatos y formas de representación, se hace imprescindible una vía de procesarla y recuperarla para que pueda ser usada con diversos fines. Por lo cual, surgen los sistemas recuperadores de información que son herramientas informáticas que permiten acceder a la información previamente almacenada (Pinto Molina 2011).

Un sistema de recuperación de información (SRI) es “un sistema de búsqueda basado en palabras claves, que incorpora automáticamente un gran número de archivos alojados en servidores web” (Belén Fuentes, 2013), están dirigidos a facilitar la navegación y recuperación de la información necesaria (Torres Pombert 2003). Los

[4] British Broadcasting Corporation
[5] Sitio informativo cubano dedicado a la lucha contra el terrorismo mediático

buscadores web o **motores de búsqueda**, son sistemas informáticos que nos dan la posibilidad de consultar una gigantesca base de datos para encontrar páginas web. Los buscadores de Internet brindan a los cibernautas la opción de **encontrar la información** que necesitan de una forma rápida, ágil y sencilla.

En Internet existen buscadores web que son usados por gran parte de los usuarios a nivel mundial, entre ellos se encuentran **Google News, Bing**, **Yahoo!, Ask.com, AltaVista** y **Aol search** que su función principal es la búsqueda de noticias (Bonilla, 2014), buscadores que son reconocidos como los más usados en el mundo según (Buscadores Web 2015)

Los buscadores antes mencionados utilizan políticas de posicionamiento particulares que afectan directamente la posición de los sitios cubanos, políticas como el posicionamiento por cantidad de visitas o cantidad de sitios que apunten a uno específico. Indicadores que dejan a los sitios cubanos en desventaja por la existencia de pocos usuarios en la red cubana, además, se incluye la posibilidad de una mala programación de las páginas cubanas al no utilizar correctamente los metadatos y normas establecidas para que los rastreadores puedan encontrar los sitios. Esto provoca que la mayoría de los sitios nacionales queden en posiciones alejadas en los resultados de búsqueda, conllevando a que los usuarios nacionales no encuentren la información pertinente que a los sitios propios del país.

Por estas razones en Cuba se han creado varios buscadores como **Lupa** que trabaja en la Red universitaria (REDUNIV) con el propósito de encontrar contenidos dispersos[6] en sitios web y repositorios públicos en la red del Ministerio de Educación Superior (MES**)** y también en las redes del Ministerio de Educación (MINED), Centro Nacional de Información en Ciencias Médicas (INFOMED), Cultura, Joven Club, Universidad de las Ciencias Informáticas (UCI) (UPREDES, 2015). Otro motor de búsqueda es **BusK2r**, perteneciente a la red de la universidad de Oriente (UO), del MES, Red Informática del Ministerio de Educación de Cuba (RIMED) e INFOMED. Todos estos buscadores tienen como objetivo solventar las necesidades de búsqueda de información de los usuarios cubanos que no poseen internet, además, de explorar sobre los contenidos propios del país.

En consecuencia al desarrollo tecnológico del país, específicamente la informatización de la sociedad, se crea el motor de búsqueda Orión con el objetivo de realizar búsquedas de contenidos propios de la red cubana. Además, de ser una contribución a la soberanía tecnológica del país junto con los demás buscadores cubanos,

[6] Se refiere a contenidos alojados en distintos sitios web de la red.

ya que se contaría con un sistema propio que respondería a las necesidades propias de los usuarios nacionales.

Actualmente cuenta con varias opciones de búsqueda para diferentes contenidos como imágenes, noticias y documentos. Por otra parte, no posee en estos momentos una forma de agrupar las noticias en una vista donde aparezca solo este tipo de información. Por esto resulta difícil encontrar información de tipo noticioso obligando al usuario a discernir una noticia de otro tipo de contenido, lo cual hace que el proceso de consumo de información noticiosa sea engorroso, lento y poco eficiente. La afirmación anterior se puede demostrar al analizar los resultados de una búsqueda de este tipo. Primero se tiene que las noticias no están en orden cronológico provocando confusión ya que en ocasiones no se observa la fecha de publicación confiando en que son las más actualizadas. También, se hace notar que en ocasiones los resultados no son los esperados, pues pueden ser arrojados entre estos, contenidos distintos al solicitado, como documentos en diferentes formatos que tienen alguna relación con el criterio de búsqueda.

Ante todos estos elementos se puede concluir que los usuarios actuales y los posibles usuarios futuros encontrarán una gran dificultad al intentar satisfacer sus necesidades de informarse del acontecer tanto nacional como internacional. Esto provocaría la disminución de usuarios del sistema, al mismo tiempo, se extendería el uso de buscadores extranjeros en los cuales los sitios cubanos en general tienen una posición desventajosa. Por otra parte, los usuarios de subredes cubanas sin acceso a internet no pueden acceder a servicios de búsqueda en internet impidiéndoles satisfacer necesidades de información que pudieran ser vitales.

A partir de los elementos expuestos, se concluye como **problema de investigación:** ¿Cómo optimizar en el buscador Orión la recuperación de noticias en el dominio .cu, para facilitar el acceso a los principales medios de prensa digitales cubanos?

El **objeto de estudio** lo enmarca el proceso de recuperación de información digital.

Como **campo de acción** se establece: el proceso de recuperación de información de noticias sobre el dominio .cu.

Con el fin de darle una solución al problema de investigación se define como **objetivo general:** Desarrollar un subsistema de búsqueda de noticias para el buscador Orión, que permita optimizar las búsquedas de noticias publicadas en el dominio .cu.

Para lograr darle cumplimiento al objetivo general se han establecido los siguientes **objetivos específicos:**

1. Caracterizar los fundamentos teóricos relacionados con la recuperación de información y el procesamiento de noticias.
2. Estudiar las tecnologías, las herramientas y la metodología para la implementación del subsistema de recuperación de noticias para el buscador Orión, con el objetivo de hacer un mejor uso de ellas.
3. Diseñar e implementar el subsistema de recuperación de noticias para el buscador Orión.
4. Validar el subsistema de recuperación de noticias para el buscador Orión.

Idea a defender:

Con la creación de un subsistema de búsqueda de noticias para Orión que permita visualizar solamente y de manera automática noticias publicadas en sitios de corte noticioso dentro del dominio cubano (cu), ayudará a los usuarios a mantenerse informados del acontecer nacional e internacional de manera sencilla, pudiendo encontrar las noticias de diversas fuentes en un mismo sitio.

Un subsistema de búsqueda de noticias para Orión permitirá recuperar las noticias bajo el dominio (cu).

Con vista a darle cumplimiento a los objetivos específicos se plantean las **siguientes tareas de investigación:**

1. Elaboración del diseño teórico - metodológico de la tesis.
2. Realización de un estudio sobre las tendencias en el desarrollo de sistemas de recuperación de información y procesamiento de noticias.
3. Selección de las tecnologías, herramientas y estándares que se necesitan para implementar la propuesta de solución y de la metodología de desarrollo de software.
4. Elaboración de los artefactos requeridos por la metodología de desarrollo seleccionada.
5. Implementación de la propuesta de solución.
6. Ejecución de las pruebas sobre la solución y documentación de las pruebas realizadas.

Del cumplimiento de las tareas propuestas se obtiene como **posible resultado:**

Un subsistema de búsqueda de noticias para el motor de búsqueda Orión que permita realizar búsquedas de noticias y visualizarlas en una misma vista de manera automática. El mismo debe permitir la funcionalidad de filtrar los resultados por fecha, medio, temática, autor, idioma o una combinación de filtros.

Para desarrollar la investigación se emplearon los siguientes **métodos científicos:**

Analítico-sintético: Empleado para el análisis de las herramientas, tecnologías y metodologías, para identificar y determinar cuales se utilizarán para el desarrollo de la investigación. Permitiendo llegar a un mayor entendimiento del objeto de estudio para aplicarlo en el desarrollo de la solución propuesta.

Análisis estático: Se utilizó para la inspección o revisión de la solución con el objetivo de encontrar errores en el sistema.

Inducción-deducción: Se utilizó en el análisis de las características de los sistemas homólogos para arribar a razonamientos que puedan ser aplicados al problema a resolver.

Modelación: Se utilizó en la representación de las características del sistema, así como las relaciones entre los objetos y funcionalidades del mismo.

Análisis documental: se utiliza en la revisión de la documentación para estudiar los conceptos asociados, informaciones relacionadas con el estudio de homólogos, o con las herramientas. Permite obtener el conocimiento necesario para el desarrollo del subsistema.

Estructura del documento:

Capítulo 1. Fundamentación teórica sobre el Subsistema de búsqueda de noticias del buscador Orión:

En este capítulo se declaran conceptos fundamentales para la comprensión de la investigación, se incluye el estudio de homólogos y se establece una comparación entre metodologías, tecnologías, herramientas y lenguajes para determinar las que serán utilizadas en el desarrollo de la solución propuesta.

Capítulo 2. Diseño de la propuesta de solución Subsistema de búsqueda de noticias del buscador Orión:

En este capítulo se definen los requisitos de software, las estructuras de datos y los artefactos necesarios para la implementación de la solución propuesta. Además, se exponen las principales características del sistema, su diseño, arquitectura y el estándar de código utilizado.

Capítulo 3. Implementación y prueba del Subsistema de búsqueda de noticias del buscador Orión:

Se muestran las vistas principales del subsistema desarrollado donde se evidencian las funcionalidades implementadas. Además, se presentan los casos de pruebas para la validación de la aplicación. Se documentan los resultados de la aplicación de los casos de prueba para validar su calidad.

Capítulo 1. Fundamentación Teórica sobre el Subsistema de búsqueda de noticias del buscador Orión.

En el presente capítulo se realiza una investigación sobre los sistemas homólogos más importantes, específicamente en el área de la recuperación de información de tipo noticia. También se brindan conceptos relacionados con el dominio del problema. Por otra parte, se explica el funcionamiento y características de las principales herramientas de búsqueda de información, y se describen las herramientas, tecnologías y lenguajes de programación que se usarán para el desarrollo de la solución.

1.1. Conceptos asociados al dominio del problema

El objetivo principal de los siguientes conceptos es obtener un conocimiento suficiente del dominio del problema como para poder comunicarse eficazmente con clientes y usuarios, comprender su negocio, entender sus necesidades y poder proponer una solución adecuada.

1.1.1. Medios digitales de comunicación

Los medios digitales de comunicación son plataformas informativas, alojadas en la web y constituidas por herramientas audiovisuales, formatos de interacción y contenidos de carácter virtual. Entre los medios digitales sobresalen los blogs o bitácoras, las revistas virtuales, las versiones digitales y audiovisuales de los medios impresos, páginas web de divulgación y difusión artística y emisoras de radio virtuales. La rapidez, la creatividad y la variedad de recursos que utilizan los medios digitales para comunicar, hacen de ellos una herramienta muy popular para su uso. Su variedad[7] es muy amplia, lo que hace que, día a día, un gran número de personas se interesen por ellos para crear, expresar, diseñar, informar y comunicar (“Banrepcultural” 2015).

1.1.2. Noticia

La noticia es la construcción periodística de un acontecimiento cuya novedad, imprevisibilidad[8] y efectos futuros sobre la sociedad lo ubican públicamente para su reconocimiento. Una noticia es la construcción de un artículo por una institución informativa con el consciente objetivo de comunicar al público un acontecimiento considerado importante y por tanto noticiable. (Bello 2012).

Noticia digital: Noticia digital o publicación digital es aquella obra cuyo formato, almacenamiento y distribución está basado en medios digitales y electrónicos, que representan hechos o acontecimientos externos o ajenos

[7] Referido a los tipos de medios digitales (periódicos, revistas, bitácoras, redes sociales)
[8] Que no se puede conocer de antemano lo que va a ocurrir.

al sujeto que la consume. Existen una gran variedad de ellas y cubre las de su versión en papel y variantes digitales: revistas, periódicos, blogs, libros, etc.(“Revistero Virtual” 2016)

1.1.3. Estándares de la Noticia digital

Los estándares son establecimiento de normas a las que debe ajustarse la información y la interoperación de los sistemas que deben manejarla. Los siguientes estándares son los más importantes a destacar (Ramos 2013):

1- Brevedad.
2- Claridad.
3- Distanciamiento y organización de la información en una pirámide invertida
    a. Información fundamental.
    b. Información secundaria.
    c. Detalles.

1.1.4. Recuperación de Información (RI)

La recuperación de la información ha desempeñado un papel esencial en la evolución de la sociedad. Hay que recordar que la recuperación de información surge para buscar soluciones y dar una respuesta al problema de la explosión de información científica; en la actualidad, la World Wide Web, como medio de acceso a la información más utilizado, juntamente con la facilidad para poder publicar en él ha provocado que uno de los principales problemas a los que se enfrenta cualquier persona es cómo localizar información pertinente ante el exceso de información existente.

Entre las definiciones encontradas en el proceso de investigación se encuentra que es el conjunto de actividades orientadas a facilitar la localización de determinados datos u objetos, y las interrelaciones que estos tienen a su vez con otros (Martínez Méndez, 2004), concepto que no esclarece demasiado pero concreta lo que realmente es. Otro concepto plantea que la RI es encontrar material, usualmente documentos, de naturaleza no estructurada que satisfaga la necesidad de información a partir de una colección enorme almacenada generalmente en computadoras (Manning, Raghavan, y Schütze 2010).

De estos conceptos podemos deducir o concluir que la RI es un conjunto de actividades como la recuperación, análisis, almacenamiento y presentación de información contenida en la web, en diferentes formatos, que es requerida por uno o varios usuarios.

1.1.5. Sistemas de recuperación de información (SRI)

EL concepto se da a conocer por Calvin Mooers en la década de 1950 (Croft 2005). Esto se evidencia en la existencia de métodos de recuperación de información en las antiguas colecciones de papiros. Otro ejemplo clásico que se ha venido utilizando es la tabla de contenidos de un libro, sustituida por otras estructuras más complejas a medida que ha crecido el volumen de información. La evolución lógica de la tabla de contenidos es el índice, estructura que aún constituye el núcleo de los SRI actuales. (BAE, 1999)

Los SRI aparecen con el propósito de manejar, la recuperación y el filtrado de la información contenida en la Web. De manera general, se pueden distinguir tres tipos de herramientas para localizar la información: Motores de búsqueda, Directorios web y Metabuscadores (McDonald 2008). Los SRI son útiles para procesar grandes cantidades de documentos rápidamente. La cantidad de datos en la red ha crecido tan rápido como la velocidad de las computadoras, y ahora utilizando los SRI somos capaces de buscar en colecciones de miles de millones de palabras contenidas en diferentes tipos de documentos.

1.1.6. Motor de Búsqueda

Son herramientas web dedicadas exclusivamente a almacenar direcciones (enlaces) de páginas relacionadas a conceptos que en determinado momento se busca (Senn 1996). Por otra parte, se tiene que es una herramienta para la búsqueda de documentos por palabras clave (criterio de búsqueda) dentro de un repositorio. Entre estas herramientas alguna requieren el uso de operadores lógicos (*and*, *or*, *not*), así también están incluidos los más avanzados que permiten el uso del lenguaje natural para realizar las búsquedas. (Cabrera Gerra, Vega Prieto, and Báez Ramos 2009)

Un motor de búsqueda suele definirse como una herramienta web que permite encontrar información cuando esta no está ordenada. A este proceso que se realiza a partir de una consulta hecha por el usuario existen varios términos asociados como *crawler* o *spiders* (Rueda and Delgado 2012). Un buscador es una herramienta que registra las páginas web según su contenido, pueden operar de manera distinta. Pueden basar los resultados en dependencia del contenido de las páginas, las etiquetas ***Meta Tag*****,** y/o el título de la página (Sunshine, n.d.).

En resumen, un buscador o motor de búsqueda es una herramienta que puede o no tener una interfaz web, que opera sobre los distintos contenidos ubicados en la red. Estos tienen como objetivo encontrar información pertinente a un criterio. Además, están compuestos generalmente por un rastreador, encargado de realizar el rastreo de los contenidos. Y un indexador que se encarga de registrar los datos sobre los contenidos

encontrados, creando un índice que permitirá encontrar la información más rápido y devolverla de acuerdo a un criterio de relevancia.

1.1.7. Crawler

"Se llama *crawling* al procedimiento de visitar páginas para ir actualizando lo que el buscador sabe de ellas". "Un ***crawler*** es un programa que corre en la máquina del buscador y que solicita a distintos computadores en Internet que le transfieran el contenido de las páginas web que él les indica." (Baeza- Yates 2008).

Un *crawler* comienza con un conjunto pequeño de páginas conocidas, dentro de las cuales encuentra enlaces a otras páginas, que agrega a la lista de las que debe visitar. Debido a esto la lista crece rápidamente y se hace necesario establecer un orden para visitar estas páginas o lo que se conoce como "política de *crawling*", para lograr esto se le atribuye una importancia a las páginas que dictaminan el orden en que actualizará el proceso de *crawling*. La importancia se puede medir a través de la cantidad de visitas que se le haga a la página o por la cantidad de páginas que le apunten (Baeza- Yates 2008).

1.1.8. Indexación

*"*El indexamiento, es el proceso de construir un índice de las páginas visitadas por el *crawler*. Este índice almacena la información, de manera que sea rápido determinar que páginas son relevantes a una consulta." (Baeza- Yates 2008).

Debido a los volúmenes de información involucrada en los procesos de búsqueda, el buscador construye un índice invertido*,* que tiene un índice de todas las palabras diferentes encontradas, almacenando una lista de las páginas donde estas palabras aparecen. Con este índice invertido la búsqueda puede resolverse buscando las palabras en el índice y resolviendo la lista de páginas relacionadas sin necesidad de realizar el proceso de búsqueda por todos los documentos (Baeza- Yates 2008).

1.2. Estudio de homólogos

Actualmente son varios los sistemas de recuperación de información que existen. Para lograr un mejor entendimiento de estos sistemas, así como para obtener una mayor aproximación a sus características y funcionalidades, además de comprender sus ventajas y desventajas. Se realiza a continuación un estudio de los principales buscadores de noticias en el ámbito internacional y nacional, se tendrán en cuenta los siguientes aspectos: rapidez, exactitud, presentación y filtros que presentan. Con el objetivo de aplicar los que sean convenientes en la solución.

1.2.1. Buscadores internacionales

Los buscadores más utilizados en los últimos años son: Google, Baidu, Bing, Yahoo!, Ask, Exite y AOL (Buscadores Web 2015). Como criterio de evaluación se usa el porcentaje de uso de los buscadores por los usuarios a nivel mundial.

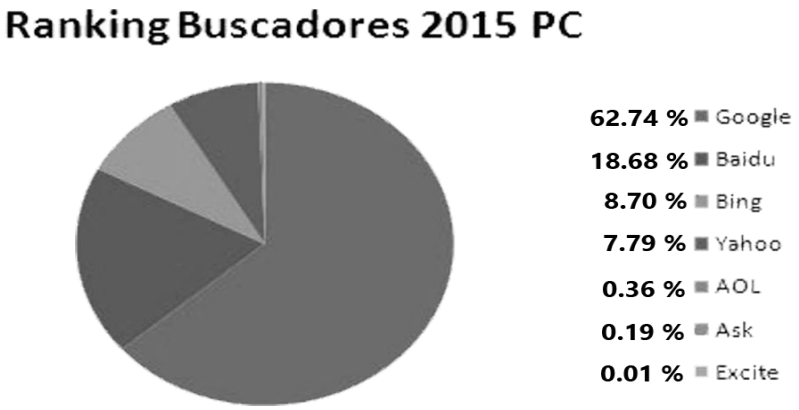


Ilustración 1: Ranking de los Buscadores para PC 2015 (Buscadores Web 2015)

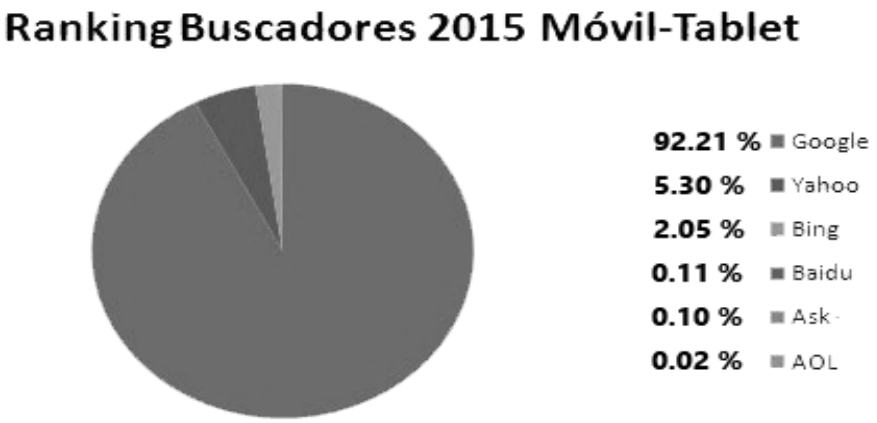


Ilustración 2: Ranking de los Buscadores para Móvil-Tablet 2015 (Buscadores Web 2015)

Google

Surgió en 1997 y más del 90 % de los usuarios de la Web lo utilizan. Este ofrece una forma simple y rápida de encontrar la información en la Web, con acceso a más de 8.168 millones de páginas web. Google posee un sitio para cada país, así como un sitio global o el principal de la empresa. Google usa varios rastreadores para

la recolección de las páginas web, el más antiguo es Googlebot, encargado de recoger los enlaces que dará como resultado en Google. (Rovira, Capdevila, and Marcos 2014)

Google Noticias

Google Noticias es un servicio lanzado por la empresa Google en el 2002, este se mantuvo en versión de prueba hasta el 2006 y hasta junio del 2012 existían ya 72 ediciones regionales para sectores geográficos concretos. El agregador de noticias está en 28 idiomas, este servicio posee un algoritmo que funciona de manera automática rastreando cada 15 minutos los más de 25.000 medios digitales de los cuales actualiza su base de datos (Rovira, Capdevila, and Marcos 2014).

La página de noticias está compuesta por tres paneles horizontales y tres verticales.

Horizontales:

1. Listado de los servicios de Google y configuración de búsqueda.
2. Identificador de Google y entrada de texto para la búsqueda, incluida la búsqueda avanzada.
3. Nombre del servicio y panel de configuración de la vista de los resultados con 4 estilos diferentes. Así como un panel para la elección de la edición que se desee.

Verticales:

1. Panel que lista las diferentes clasificaciones de las noticias y un listado de las más destacadas.
2. Bloque para mostrar las noticias según su clasificación apareciendo por el mismo orden de clasificación que en el primer panel.
3. Panel de las noticias más populares. Las que más se visitan.

Por cada resultado que este servicio muestra se puede apreciar diferentes elementos incluidos. (Rovira, Capdevila, and Marcos 2014)

En el bloque donde se muestran las noticias estas presentan 5 elementos distinguibles como es el **título** que describe a grandes rasgos el tema de la noticia, dando una idea general sobre el tema. También es un enlace activo que lleva al artículo completo en la fuente original. Como segundo elemento aparece una **imagen** tomada a partir del artículo original. Luego el **resumen**, este es una muestra de la noticia de entre las primeras 26 a 34 palabras, para representar en pocas palabras el contenido de esta. Después está la **fuente** que denota el nombre del sitio de donde se toma la noticia o sea el medio en el que se encontró. Por último, la **fecha** que

informa sobre el momento en que se publicó la noticia. Esta fecha tiene dos formatos, primero, Día\Mes\Año y segundo si la fecha es anterior a la actual muestra el tiempo desde que se publicó la noticia si es de la fecha actual dígase hora(s) y minuto(s).

Cuando se encuentran noticias similares a la actual entonces se muestran varios campos adicionales en la noticia. Estos son título de la noticia secundaria, fuente secundaria, listado de noticias relacionadas con el tema y por último aparecen otros dos campos, imagen relacionada con la noticia y fuente de donde se extrajo la imagen.

Baidu

Baidu con un surgimiento reciente es el buscador más popular de China, lo que lo hace un poderoso canal para los negocios en el mercado. Al final del 2013 tenía en su poder el 63.55 % de la cuota de mercado de los motores de búsqueda en China. Para cualquier compañía de negocios que tenga un sitio web chino, tener una optimización por Baidu se ha convertido en una necesidad. (SearchEngineLand 2015b)

Este tiene una optimización única muy parecida a la de Google 2009. La optimización consiste en crear muchos contenidos únicos de alta calidad y cantidad de contenido, cumpliendo con requisitos técnicos específicos para Baidu. Entre los requerimientos se encuentran (SearchEngineLand 2015b): No duplicar contenido, Baidu penaliza los sitios web más que otros motores de búsqueda por el duplicado de contenido, asegurando que cada parte de contenido del sitio sea única; lenguaje de uso, todo el contenido y los metadatos deben estar escritos en caracteres chinos simplificados; definiciones de contenido, en el idioma chino, existen múltiples dialectos y múltiples significados para una sola palabra. Se deben comprobar todas las palabras del contenido para asegurar que se están usando las palabras correctas; etiquetas del título, las etiquetas del título deben ser escritos de la misma manera que lo haría para Google.

Bing

Bing es una herramienta o buscador web perteneciente a Microsoft[9], su puesta en servicio fue el 3 de junio de 2009. Este tiene por objetivo facilitar las búsquedas, haciéndolas rápidamente y mostrando el resultado lo más preciso posible. Tiene como característica cambiar todos los días la imagen de fondo en el buscador las cuales corresponden a lugares importantes del mundo. Otra característica es que se pueden dividir los resultados según su tipo, además incluye un panel de navegación así como información sobre búsquedas relacionadas.(Cabrera and Castillo Pérez 2013)

[9] Microsoft Corporation es una empresa multinacional de origen estadounidense dedicada al sector del software y el hardware

Ofrece la posibilidad de hacer uso de servicios como Imágenes, Noticias, Videos y Mapas. Este buscador también ofrece ventajas importantes como la obtención rápida de resultados, filtro de contenido adulto activado por defecto y un historial de búsqueda. La vista de noticias de Bing cuenta con tres paneles horizontales donde se encuentran el buscador como tal, varios elementos de configuración, una lista de enlaces que llevan a los demás servicios de Bing y los filtros disponibles los cuales son el filtro de Fecha y el de coincidencia.

Yahoo

Es un caso muy especial porque debe ser el portal de búsqueda de Internet más viejo y conocido, pero la mayoría de sus usuarios no saben que Yahoo! es principalmente un directorio Web y no un verdadero Motor de Búsqueda. Por lo tanto, su estructura está basada en sitios web propuestos por usuarios y no por los encontrados por un Robot o Spider. Creado por David Filo y Jerry Yang, Ingenieros Eléctricos de la Universidad de Stanford, comenzaron su catálogo en abril de 1994, para mantener y rastrear los sitios de su propio interés. El Motor de Búsqueda de Yahoo fue Google hasta febrero de 2004. Actualmente Yahoo usa su propio motor de búsqueda el cual está basado en Inktomi (buscador que adquirió en marzo de 2003) y basado también en AltaVista y en AllTheWeb. (Dolores 2011)

El año 2011 fue un año clave para Yahoo, para ese tiempo, si no antes este planeaba renunciar a su tecnología de búsqueda. en el trato entre Microsoft y Yahoo, este mantiene todavía su propia marca. El buscador puede alterar su aspecto y la sensación entorno a los resultados, así como proporcionar información adicional. Lo que lo hace más conveniente como motor de búsqueda. Además de la búsqueda web también posee varios servicios. Entre los que se encuentran Yahoo Answers, Yahoo`s Delicious, Yahoo`s Flickr, Yahoo Search, Yahoo Seo y Yahoo News. (SearchEngineLand 2015a)

Yahoo Noticias

Este servicio se presenta como una página web compuesta por tres paneles horizontales donde se encuentran en el orden que se mencionan, una lista de enlaces que llevan a los demás servicios disponibles, el buscador como tal junto con la opción de inicio de sesión y el acceso a correo, por último, los filtros disponibles que son: política, mundo, tecnología, ciencia, virales, coches y videos.

La sección correspondiente a las noticias en sí posee dos paneles verticales donde se muestran las noticias. En el panel izquierdo tiene un bloque donde se presenta la primera noticia mostrando primero la imagen correspondiente con las dimensiones del bloque seguido del título, resumen, fecha y autor todo en la parte

inferior izquierda utilizando el 50 % del bloque, dejando el resto para presentar una secuencia de tres títulos correspondientes a las publicaciones más destacadas junto con la fecha de publicación.

A continuación de este bloque se presenta un listado del resto de las noticias utilizando un formato de título, fecha, resumen e imagen en la parte izquierda de los anteriores. El título describe a grandes rasgos el tema de la noticia, dando una idea general sobre el tema. También es un enlace a otra página de Yahoo donde se muestra la noticia en su totalidad. La fecha informa sobre el momento en que se publicó la noticia, en este caso solo muestra el día y mes correspondiente. Seguido se encuentra el resumen, este es una muestra de la noticia para representar en pocas palabras el contenido. También incluye una imagen tomada a partir del artículo original encima de la fuente de donde se toma el artículo.

En el bloque derecho aparecen distintas publicaciones que no corresponden a un tema específico, es decir, que pueden corresponderse a distintas temáticas como salud, cultura o alguna curiosidad. Estas tienen un formato compuesto por imagen, título y autor.

Por último, en la vista correspondiente a la noticia completa se muestra el título que ya no es un enlace, la fuente, fecha completa, imagen y cuerpo de la noticia que es el texto donde se desarrolla la descripción del hecho. Además, cuenta con varias opciones debajo como compartir y enviar por correo, al final se muestran comentarios sobre la noticia.

Aol search

AOL (América Online) es uno de los portales más visitados en Estados Unidos y cuenta con páginas en distintos idiomas que se adaptan a cada país. Todo el sistema de clasificación de páginas web está administrado por Google™ ya que son un mismo grupo empresarial.

El buscador de AOL (AOL Search) hace muy simple encontrar lo que se necesita. También permite descubrir todo tipo de información, incluyendo videos, fotos, sonidos y noticias. Este posee un buscador de imágenes que ordena los resultados por relevancia. La búsqueda de video, incluye varios formatos (Real Media, Windows Media, Quick Time, Mp3 y Flash), se realiza sobre los contenidos de AOL y de toda la red ordenados por relevancia. (SlideShare 2015)**.** También es posible realizar la búsqueda de noticias, gracias a una importante base de datos que incluye artículos y contenido de varios proveedores como la asociación de prensa, Reuters, CNN y revista como People y Fortune.

Presenta navegación por pestañas, cuenta con un motor de navegación muy eficaz. Es capaz de trabajar con diferentes protocolos. Presenta un aspecto visual llamativo y tiene un alto nivel de personalización. Aol está compuesto por distintos bloques. El bloque de búsqueda donde aparece el buscador, el bloque para la selección del tipo de búsqueda (en la Web o imágenes). Un bloque que contiene dos listados. El primero con las búsquedas relacionadas con el tema y el segundo con el historial de búsqueda. El bloque de resultados está dividido en tres partes la primera y la última corresponde a publicidad, dejando por el medio el de las noticias. Las que se muestran utilizando los campos título, fuente y resumen.

Ask

Es parte de la compañía InterActive Corporation, fundada en 1996 por Garrett Gruener y David Warthen en Berkeley, California. El programa original fue implementado por Gary Chevsky basado en su propio diseño. Los primeros inversionistas fueron el Grupo RODA. La idea que dio lugar a Ask.com fue la capacidad de responder preguntas realizadas en lenguaje natural. Ask.com fue el primer buscador comercial del tipo pregunta-respuesta desarrollado para www.

Permite una amplia variedad de consultas realizadas en inglés, así como las tradicionales búsquedas con palabras clave, y pretende que las búsquedas sean más intuitivas que con los otros buscadores. Ask Jeeves vendió la misma tecnología que utiliza en el sitio Ask.com a distintas empresas como Dell, Toshiba y E*Trade. Parte de la empresa fue vendida a Kanisa en el año 2002. Desde la compra del motor de búsqueda Teoma por Ask.com en 2001, los buscadores de esta compañía utilizan una tecnología denominada ExpertRank. Al contrario de lo que ocurre con el PageRank que utiliza Google, con ExpertRank, los enlaces a un sitio web tienen más peso si proceden de otros sitios dedicados a un tema similar (Dolores 2011).

Para acceder a las noticias de Ask se debe hacer una búsqueda en la página principal sobre el tema que se desea. Esta lleva a una vista donde se despliegan distintos artículos relacionados con el criterio de búsqueda. Para acceder a las noticias se necesita acceder a través del vínculo **news** que aparece en la parte superior, antecedido por el buscador y otros enlaces de acceso. Las noticias se muestran en la parte izquierda presentando primeramente un grupo de imágenes relacionadas con el tema. Seguido de las noticias con un formato compuesto por el título, imagen, resumen, número de artículos del mismo tema, las principales fuentes de dichos artículos, número de imágenes sobre la noticia, fuente de la noticia que se muestra y fecha.

Excite

Excite Italia nace en mayo de 1999 y es, actualmente, un *network* europeo activo en Italia, Reino Unido, Alemania, Francia, España, Holanda y Austria. Cuenta en total con 5 millones aproximadamente de visitantes únicos al mes. En los años 2003 - 2004 Excite vuelve a abrir propiedades en los principales países europeos, relanzando excite.co.uk, excite.fr, excite.es, excite.de y excite.nl. Excite es el único portal privado que conjuga las características únicas e individuales de sitios personales con los contenidos, los instrumentos de búsqueda y comunicación de los portales tradicionales. Objetivo principal de Excite es ofrecer contenidos y servicios a medida, desarrollándolos en base a las exigencias y preferencias de cada usuario. (Exite 2015)

Para acceder a las noticias de Exite.com al entrar al sitio se observa un panel en el centro con la opción de explorar el sitio, es este se encuentra el enlace **news** que lleva a la vista de noticias. Donde estas se muestran divididas por categorías como más destacadas, nacionales, tecnología, entretenimiento, del mundo y noticias extrañas. Estas noticias se presentan mostrando el título, imagen, fuente, resumen y fecha al lado de la categoría que las agrupa. Estos elementos cumplen con las descripciones de los anteriores descritos.

1.2.2. Buscadores Nacionales.

La Web cubana cuenta con varios buscadores como **Red Cuba, Lupa y Busk2r.** Pero estos no cuentan con un subsistema de búsqueda de noticias. Actualmente la búsqueda de noticias se realiza de manera muy compleja. Ya que hay que saber bien que es lo que se busca y entre los resultados arrojados muchos de ellos no son noticias.

Plataforma de servicios CUBA (CUBA 2015)

CUBA (por sus siglas en español, Contenidos Unificados para Búsqueda Avanzada) es una plataforma que responde a las búsquedas realizadas por los usuarios y emite un listado de los sitios que coinciden con los criterios introducidos. La plataforma brinda acceso a los contenidos publicados en la red cubana, logrando que dicha información sea accedida por los usuarios.

CUBA, permite consultar información de cualquier tema que se encuentre en la red cubana de una forma más sencilla. Es capaz de acceder a la información resultante de eventos científicos, materiales de investigación, materiales educativos y otros de igual valor e importancia para la audiencia que lo solicita. Puede encontrar información en sitios, libros y tesis desarrollados por autores cubanos, así como encontrar imágenes en la red de Cuba.

En cuanto a la búsqueda de noticias, CUBA no cuenta con un subsistema de búsqueda de noticias por lo que se necesita hacerla como si se estuviera buscando normalmente algún contenido. Esta muestra los resultados presentando el título de las noticias, que puede o no coincidir con lo solicitado, la fuente de origen de las noticias y por último un conjunto de títulos que corresponden al resto de las noticias encontradas en la fuente, estos aparecen sin ningún delimitador que defina donde comienza y acaba un título lo que lo hace confuso para diferenciar entre títulos.

Lupa

Es un buscador de contenidos en la REDUNIV. Con el propósito de encontrar contenidos dispersos en sitios web y repositorios públicos en la red del MES y también en las redes MINED, INFOMED, Cultura, Joven Club, UCI (UPREDES 2015). Al realizar una búsqueda la vista de los resultados se divide en dos bloques. El primero en la parte izquierda muestra el logo y nombre del buscador. Seguido ve un mensaje de bienvenida por defecto, en el que se anuncia las ubicaciones donde opera. Luego la caja de búsquedas para introducir los criterios de búsqueda. Además, lupa es capaz de organizar por tipo de archivo los resultados obtenidos en una consulta gracias a una serie de filtros para acotar los resultados. Por último, tiene una serie de búsquedas relacionadas que se utilizan para ayudar al usuario a direccionar su búsqueda hacia los resultados que desea.

En este buscador las noticias se muestran en el panel derecho. Donde se muestran el título y la fuente de la noticia. En caso de ser una noticia el resultado, puesto que este está enfocado principalmente en búsqueda académica. Por lo que es necesario conocer prácticamente el título de lo que se busca, esto hace que no sea una opción adecuada para buscar este tipo de contenido.

Busk2r

Este es desarrollado en la UO con el objetivo de operar sobre la red de la UO[10], el MES[11], RIMED e INFOMED. Aquí las noticias se muestran en una columna. Donde se muestra el título que es un enlace directo al origen de la noticia. Se muestra un breve resumen de la noticia y la fuente de procedencia. Con este buscador se puede elegir entre la red MES y la red UO para realizar las búsquedas esto quiere decir que se puede elegir buscar en la red del MES o en la red propia donde se aloja el buscador en la Universidad de Oriente.

En la tabla siguiente se realiza una comparación simple entre algunos sistemas homólogos tanto nacionales como internacionales, incluyendo el sistema para el que se pretende desarrollar la solución. Con el objetivo de

[10] Universidad de Oriente
[11] Ministerio de Educación Superior

demostrar las deficiencias del sistema actual en cuanto la búsqueda de noticias y por tanto la necesidad e importancia del subsistema a desarrollar.

Tabla 1: Comparación entre sistemas Homólogos

Elementos/Buscador	Google	Bing	Yahoo	C.U.B.A	Lupa	Orión
Tipo	Si	Si	Si	No	No	No
Título	Si	Si	Si	Si	Si	Si
Imagen	Si	Si	Si	No	No	No
Fuente	Si	Si	Si	Si	Si	Si
Fecha	Si	Si	Si	No	No	No
Resumen	Si	Si	Si	No	No	No
Autor	Si	Si	Si	No	Si	No

1.2.3. Resultados del estudio de los sistemas homólogos.

A partir del estudio de los sistemas homólogos se demuestra la necesidad de adicionar funcionalidades ya existentes en buscadores a nivel internacional y no a nivel nacional, a la solución propuesta para brindar al usuario facilidades en búsquedas de noticias haciendo uso del buscador cubano Orión.

1.3. Tecnologías

Para desarrollar el subsistema de búsqueda de noticias para el buscador Orión que se propone, se hace necesario investigar sobre las tecnologías a utilizar. A continuación, se procede con el estudio y selección de las mismas.

1.3.1. Indexadores

El indexador tiene como propósito la elaboración de un índice que contenga de forma ordenada la información, esto con la finalidad de obtener resultados de forma sustancialmente más rápida y relevante al momento de realizar una búsqueda. (Baeza- Yates 2008)

Elasticsearch 1.0

Elasticsearch es un motor de búsqueda de código abierto construido sobre Apache Lucene[12] desarrollado en el lenguaje de programación Java y está publicado como código abierto. Es un producto que permite indexar y analizar casi al instante grandes cantidades de datos. Además, es escalable a cientos de servidores, así como a grandes cantidades de datos. (“Elastic” 2016)

[12] Lucene es una biblioteca de búsqueda de texto completo en Java de código abierto que hace que sea fácil añadir funcionalidad de búsqueda a una aplicación o página web.

Además, es un motor de búsqueda distribuido que permite el análisis casi en tiempo real. Esto le permite explorar los datos a una velocidad y en una escala nunca antes vista. Se utiliza para la búsqueda de texto completo, búsqueda estructurada, análisis, y los tres en combinación. (elastic.co 2016) Elasticsearch se puede ejecutar en computadoras o servidores de bajas prestaciones, así como escalar a cientos de servidores y petabytes de datos. Además, se puede descargar, utilizar y modificar de forma gratuita ya que está disponible bajo la licencia Apache 2, una de las licencias de código abierto más flexibles.

Sphinx

Es un motor de búsqueda de texto completo, de código abierto y libre. Está escrito en C++ y funciona en Linux (RedHat, Ubuntu, etc.), Windows, MacOS, Solaris, FreeBSD, y algunos otros sistemas (Nugraha 2014).

Es un motor de búsqueda abierto diseñado con el fin de indexar contenidos de bases de datos. Actualmente soporta de manera nativa MySQL, PostgreSQL y bases de datos ODBC. Otras fuentes de datos pueden ser indexadas mediante el apropiado filtro XML. Se distribuye en los términos GPLv2 de GNU o con licencia privativa.

Entre sus principales características se encuentra su gran escalabilidad probada hasta miles de millones de documentos, terabytes de datos y miles de consultas por segundo. Alta velocidad de búsqueda (hasta 150-250 consultas / seg por núcleo contra 1.000.000 de documentos, 1,2 GB de datos). Además, soporta booleanos, frases, proximidad por palabras y otros tipos de consultas.

Solr

Es una plataforma de búsquedas basada en Apache Lucene, esta funciona como un “servidor de búsquedas”. Entre sus principales características se incluye la búsqueda de texto completo, resaltado de resultados y manejo de documentos. Solr es escalable, permitiendo realizar búsquedas distribuidas. Solr está escrito en Java e incluye muchas características avanzadas. La principal característica de Solr es su API tipo REST[13] ya que se pueden hacer peticiones HTTP y obtener resultados XML. (Lucene.apache.org 2015) Solr utiliza la librería de Java Lucene, para la indexación de texto completo y de búsqueda (Medina García, Martínez Chong, and Hidalgo Delgado 2011).

La replicación es independiente del sistema operativo, para que pueda reproducir el mismo índice en una plataforma Windows, así como en Linux. Los índices se pueden configurar para operar en sincronía: Si alguno

[13] REST: Representational State Transfer-Transferencia de Estado Representacional

de los índices se modifica, las demás copias de dicho índice se actualizan automáticamente. Ofrece un API REST y un API de Java. Permite importar datos desde una base de datos. Distribuido bajo la licencia de Apache 2.

Selección del indexador

Partiendo de los elementos expuestos anteriormente se concluye que se utilizará Solr. Para esta elección se tuvo en cuenta, que Solr es escalable permitiendo búsquedas distribuidas. También se tuvo en cuenta el lenguaje en que está desarrollado (Java), lo que permitirá incrementar sus funcionalidades de ser necesario. Otro elemento es que esté indexador es el usado en el motor de búsqueda Orión, lo cual permitirá una mejor integración con el buscador.

1.3.2. Rastreador

Un rastreador web es un programa diseñado para navegar por la red y, de manera sistemática y organizada, indexar el contenido de las páginas web que encuentra. Los rastreadores están programados para funcionar automáticamente, siguiendo los enlaces que va encontrando en las páginas web.

Wire

Es un proyecto iniciado por el Centro de Investigación Web de Chile, diseñado como una herramienta para la recuperación de información en la Web (“WIRE” 2015)

Actualmente incluye un formato simple para el almacenamiento de colecciones de documentos Web, un rastreador Web y varias herramientas para la extracción de estadísticas de la colección Web así como herramientas para la generación de reportes sobre la colección Web.

Presenta una gran escalabilidad puesto que está diseñado para trabajar sobre grandes cantidades de documentos. Es desarrollado en C/C++ para un alto rendimiento. Todos los parámetros para el rastreo y la indexación se pueden configurar a través de un archivo XML. Incluye varias herramientas para analizar, extraer estadísticas y la generación de informes sobre subconjuntos de la Web, por ejemplo: la Web de un país o de una gran intranet. Su código esta libremente disponible (Department of Computer Sciences University of Chile, 2011).

Heritrix

Es un rastreador de código abierto y extensible a gran escala, utilizado en Internet Archive, desarrollado en Java. Este buscador busca recolectar y preservar los artefactos digitales de la cultura, en beneficio de los investigadores y las futuras generaciones. (“Heritrix - Home Page” 2016)

Entre otras características incluye su capacidad para correr varios rastreadores simultáneamente. Un único archivo de configuración en XML basado en Spring y una extensibilidad mejorada a través del Framework de Spring.

Nutch 1.11

Nutch es un *Web Crawler* libre y de código abierto desarrollado en Java bajo la licencia de Apache. Este proporciona interfaces extensibles para implementaciones personalizadas. Inicialmente fue implementado sobre la base de Apache Lucene, aunque ya la versión actual es independiente de Lucene. La arquitectura de Nutch es muy flexible permitiendo realizarle mejoras por parte de los usuarios a través de *plugins*. Este es independiente del servidor de indización lo que permite la integración con Solr. (NutchWiki 2015)

Nutch ofrece una solución transparente, pues al ser una tecnología de código abierto es posible conocer como organiza el ranking de resultados de las búsquedas. Está desarrollada en Java, y basa su arquitectura en la plataforma Hadoop de desarrollo de sistemas distribuidos.

Este buscador no distingue entre mayúsculas y minúsculas. Usando comillas (") al principio y al final de un grupo de palabras o frase realiza la búsqueda de ese texto exacto. Añadiendo el signo más (+) delante de una palabra fuerza la búsqueda de palabras no habituales. Añadiendo el signo menos (-) delante de una palabra realiza la búsqueda excluyendo esa palabra.

Selección del rastreador

Partiendo de los elementos antes expuestos se concluye que el rastreador a utilizar será Nutch debido a la flexibilidad de su estructura. Además, que permite el añadido de *plugins*. Y es dinámicamente escalable y tolerante a fallos además que es estable. Otro elemento es que es el mecanismo de rastreo utilizado por el buscador cubano Orión, lo que facilitará la integración entre ambos sistemas.

1.3.3. Marco de trabajo para PHP

Los marcos de trabajo definen un conjunto de objetos y conceptos estandarizados que permiten ser reutilizados en el diseño y desarrollo de nuevos sistemas de información.

CodeIgniter

Es un marco de trabajo de código abierto para aplicaciones web en PHP. CodeIgniter tiene muchas características que lo hacen destacar entre sus homólogos. Ya que a diferencia de otros *Frameworks* es muy exhaustivo con la documentación. En el lado de la programación es compatible con PHP4 y PHP5, por lo que es posible que corra en la mayoría de los servidores Web. También usa el patrón Modelo Vista Controlador. Lo que permite organizar la aplicación en diferentes partes como son el modelo, la vista, y el controlador. CodeIgniter también tiene una implementación del patrón Active Record, lo que hace sencillo escribir consultas SQ y hace la aplicación más entendible.

Zend Framework 2

Es un framework de código abierto para el desarrollo de aplicaciones Web y servicios Web usando PHP 5.3 o superior. Este usa solamente código orientado a objetos y utiliza la mayoría de las nuevas funcionalidades del lenguaje. La estructura es única, cada componente está diseñado con pocas dependencias a otros componentes. También ofrece una implementación robusta y de alto rendimiento utilizando el patrón arquitectónico Modelo Vista Controlador (MVC).

Symfony 2.7

Es un framework completo diseñado para optimizar, gracias a sus características, el desarrollo de las aplicaciones Web. Este separa la lógica de negocio, la lógica de servidor y la presentación de la aplicación. Proporciona varias herramientas y clases encaminadas a reducir el tiempo de desarrollo de una aplicación. Además, automatiza tareas comunes, permitiendo al desarrollador dedicarse completamente a los elementos específicos de la aplicación. Symfony está desarrollado completamente en PHP5. También es multiplataforma, pudiéndose ejecutar en plataformas tales como Windows.

Framework de fácil instalación y configuración, independiente del sistema gestor de bases de datos. Sencillo de usar y flexible. Solo se debe configurar lo que no es convencional. Sigue la mayoría de las buenas prácticas y patrones Web.

Selección del marco de trabajo para PHP.

Partiendo de las características expuestas anteriormente se concluye que el *framework* a utilizar será Symfony 2.7. Para esta selección se tuvo en cuenta que es rápido, adaptable, flexible y fácil de instalar. Además, que es el utilizado para el desarrollo del Motor de Búsqueda Orión.

1.3.4. Marco de trabajo para CSS3 y Java Script

Bootstrap 3.3.5

Es un framework o conjunto de herramientas de software libre para diseño de sitios y aplicaciones Web. Contiene plantillas de diseño con tipografía, formularios, botones, cuadros, menús de navegación y otros elementos de diseño basado en HTML y CSS, así como, extensiones de JavaScript opcionales adicionales.

Bootstrap es modular y consiste esencialmente en una serie de hojas de estilo LESS que implementan la variedad de componentes de la herramienta. Una hoja de estilo llamada Bootstrap. Less incluye los componentes de las hojas de estilo. Los desarrolladores pueden adaptar el mismo archivo de Bootstrap, seleccionando los componentes que deseen usar en su proyecto.

Para todos los niveles de habilidad presenta un diseño adaptativo, plugins jQuery personalizados, presenta un soporte para todos los principales navegadores y para Tabletas y Teléfonos inteligentes también. Es una de las herramientas de front-end (tecnologías que corren del lado del cliente) más completas con docenas de componentes totalmente funcionales.(Mark Otto and Jacob Thornton 2016)

Blueprint

Es un *framework* CSS, cuyo objetivo es reducir su tiempo de desarrollo. Se le da una base sólida para construir su proyecto en la parte superior de, con una rejilla de fácil uso, tipografía sensible, plugins útiles, e incluso una hoja de estilo para la impresión. Un restablecimiento CSS que elimina las discrepancias a través de los navegadores. Una red sólida que puede soportar el más complejo de los diseños. Tipografía basada en los principios de expertos que son anteriores a la Web.

Selección del marco de trabajo para CSS3 y JS.

Teniendo en cuenta los elementos antes expuestos se decide utilizar el *framework* Bootstrap. Debido a que es libre y de código abierto. Además de ser potente, también incluye plugins jQuery y da soporte a la mayoría de

los buscadores del mundo. Otro elemento es que es el usado en el desarrollo del motor de búsqueda Orión. Lo que hará sencillo la integración, así como cualquier modificación posterior.

1.4. Lenguajes empleados

Un lenguaje de programación es un lenguaje diseñado para describir el conjunto de acciones consecutivas que un equipo debe ejecutar. Por lo tanto, un lenguaje de programación es un modo práctico para que los seres humanos puedan dar instrucciones a un equipo. (Creative Commons, 2014)

1.4.1. Lenguajes de programación del lado del servidor

Existe una multitud de lenguajes concebidos o no para Internet. Cada uno de ellos explota más a fondo ciertas características que lo hacen más o menos útiles para desarrollar distintas aplicaciones. Un lenguaje del lado del servidor es aquel que se ejecuta en el servidor web, justo antes de que se envíe la página a través de Internet al cliente. Las páginas que se ejecutan en el servidor pueden realizar accesos a bases de datos, conexiones en red, y otras tareas para crear la página final que verá el cliente. Los lenguajes de lado servidor más ampliamente utilizados para el desarrollo de páginas dinámicas son el ASP, JSP, PERL y PHP

PHP

Es un lenguaje de 'scripting' de propósito general y de código abierto que está especialmente pensado para el desarrollo Web y que puede ser embebido en páginas HTML. Su sintaxis recurre a C, Java y Perl, y es fácil de aprender. La meta principal de este lenguaje es permitir a los desarrolladores Web escribir dinámica y rápidamente páginas web generadas; aunque se puede hacer mucho más con PHP. Puede usarse en todos los principales sistemas operativos, incluyendo Linux, muchas variantes de Unix (incluyendo HP-UX, Solaris y OpenBSD), Microsoft Windows, Mac OS X, RISC OS. PHP admite la mayoría de servidores Web de hoy en día, incluyendo Apache, IIS, y muchos otros. Esto incluye cualquier servidor Web que pueda utilizar el binario de PHP FastCGI, como lighttpd y nginx.(php.net 2016)

Como principales ventajas se encuentra que es un lenguaje muy fácil de aprender. Se caracteriza por ser un lenguaje muy rápido, el cual soporta la programación orientada a objetos. clases y herencia. Es un lenguaje multiplataforma. Presenta una capacidad de conexión con la mayoría de los manejadores de base de datos: MySQL, PostgreSQL, Oracle, MS SQL Server, entre otras. Capacidad de expandir su potencial utilizando módulos. Posee documentación en su página oficial la cual incluye descripción y ejemplos de cada una de sus

funciones. Es libre, por lo que se presenta como una alternativa de fácil acceso para todos. Incluye gran cantidad de funciones. No requiere definición de tipos de variables ni manejo detallado del bajo nivel.

ASP.NET

Este es un lenguaje comercializado por Microsoft, y usado por programadores para desarrollar sitios web. ASP.NET es el sucesor de la tecnología ASP, fue lanzada al mercado mediante una estrategia de mercado denominada .NET. El ASP.NET fue desarrollado para resolver las limitantes que brindaba su antecesor ASP. Creado para desarrollar web sencillas o grandes aplicaciones. Para el desarrollo de ASP.NET se puede utilizar C#, VB.NET o J#. Los archivos cuentan con la extensión (aspx). Para su funcionamiento de las páginas se necesita tener instalado IIS con el Framework .Net. Microsoft Windows 2003 incluye este framework, solo se necesitará instalarlo en versiones anteriores.(Microsoft 2016)

Ventajas de este lenguaje: completamente orientado a objetos. División entre la capa de aplicación o diseño y el código, lo cual facilita el mantenimiento de grandes aplicaciones. Incremento de velocidad de respuesta del servidor, lo que genera una mayor velocidad y seguridad. Como principales ventajas se encuentra que es privativo y presenta un mayor consumo de recursos.

Python

Es un lenguaje de programación creado en el año 1990 por Guido van Rossum, es el sucesor del lenguaje de programación ABC. Python es comparado habitualmente con Perl. Los usuarios lo consideran como un lenguaje más limpio para programar. Permite la creación de todo tipo de programas incluyendo los sitios web. Es un lenguaje de programación multiparadigma, lo cual fuerza a que los programadores adopten por un estilo de programación particular.

Este lenguaje es libre y de fuente abierta. Lenguaje de propósito general. Presenta gran cantidad de funciones y librerías. Es muy sencillo y rápido de programar, multiplataforma, orientado a objetos y portable. A pesar de todo esto presenta algo de lentitud por ser un lenguaje interpretado.(python.org 2016)

Selección del lenguaje del lado del servidor a utilizar

Teniendo en cuenta que PHP es un lenguaje multiplataforma que puede operar en distintos sistemas operativos como Windows y Linux, caracterizarse por ser un lenguaje rápido, de fácil aprendizaje y poseer la

capacidad de expandir su potencial utilizando módulos se decide optar por PHP como lenguaje del lado del servidor.

1.4.2. Otros lenguajes usados

Lenguajes de lado servidor que son aquellos lenguajes que son reconocidos, ejecutados e interpretados por el propio servidor y que se envían al cliente en un formato comprensible para él. Estos son necesarios para el desarrollo de la aplicación.

Java

Es un lenguaje de programación de propósito general, concurrente, orientado a objetos que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible. Su intención es permitir que los desarrolladores de aplicaciones escriban el programa una vez y lo ejecuten en cualquier dispositivo (conocido en inglés como *WORA*, o "*write once, run anywhere*"), lo que quiere decir que el código que es ejecutado en una plataforma no tiene que ser recompilado para correr en otra.(“Conozca Más Sobre La Tecnología Java” 2016)

HTML

Es el lenguaje con el que se escriben las páginas web. Las páginas web pueden ser vistas por el usuario mediante un tipo de aplicación llamada navegador. Podemos decir por lo tanto que el HTML es el lenguaje usado por los navegadores para mostrar las páginas web al usuario, siendo hoy en día la interfaz más extendida en la red.(MDN 2016)

Twig

Twig es a la vez, un amigable ambiente para el diseñador y desarrollador apegado a los principios de PHP, añadiendo útiles funcionalidades a los entornos de plantillas. Las características claves se encuentran en su rapidez: Twig compila las plantillas hasta código PHP regular optimizado. El costo general en comparación con código PHP regular se ha reducido al mínimo. Es considerado como seguro: Twig tiene un modo de recinto de seguridad para evaluar el código de plantilla que no es confiable. Esto te permite utilizar Twig como un lenguaje de plantillas para aplicaciones donde los usuarios pueden modificar el diseño de la plantilla. Además de ser flexible: Twig es alimentado por flexibles analizadores léxico y sintáctico. Esto permite al desarrollador definir sus propias etiquetas y filtros personalizados, y crear su propio DSL.(TwigHomepage 2016)

YAML

Los archivos de configuración de Symfony2 se pueden escribir en PHP, en XML o en YAML. Desde el punto de vista del rendimiento no hay ninguna diferencia entre los tres, ya que todos ellos se transforman a PHP antes de ejecutar la aplicación. YAML es probablemente el formato más equilibrado, ya que es mucho más conciso que XML y es bastante flexible. Su gran desventaja es que no se puede validar automáticamente, por lo que la mayoría de los errores, sólo se pueden descubrir al ejecutar la aplicación.(“YAML Ain’t Markup Language” 2016)

1.5. Herramientas utilizadas

A continuación, se detallan herramientas usadas para el desarrollo de la aplicación web y posteriores pruebas.

1.5.1. NetBeans 8.0

Es un proyecto de código abierto dedicado a proporcionar un sólido desarrollo de software. Enfocado en las necesidades de los desarrolladores, usuarios y las empresas que dependen de NetBeans como base para sus productos; en particular, para que puedan desarrollar estos productos de forma rápida, eficaz y sencilla. (“An Introduction to NetBeans” 2015). NetBeans IDE es una herramienta de desarrollo modular para una amplia gama de tecnologías de desarrollo de aplicaciones. El IDE de base incluye un editor avanzado múlti-idioma, depurador y perfiles, así como herramientas de control de versiones y colaboración desarrollador. (“NetBeans IDE - Base IDE Features” 2015)

1.5.2. Servidor Web (Apache)

Se realizó una comparación entre los servidores web Nginx y Apache para seleccionar el que más se adecuara dependiendo de las necesidades del desarrollo del subsistema. A partir de esta comparación se descartó el uso de Nginx pues la configuración es compleja y varia en dependencia del tipo de framework usado, además de que el mismo necesita de la instalación de módulos para reconocer el lenguaje php, lenguaje usado en el desarrollo de la aplicación web. Por su parte la instalación del servidor Apache es sencilla y fácil de configurar, trae por defecto reconocer el lenguaje php además de ser multiplataforma y de código libre. Es altamente configurable y de diseño modular, por lo que es muy sencillo ampliar sus capacidades. Apache permite personalizar la respuesta ante los posibles errores que puedan suceder en el servidor y es el utilizado por los desarrolladores del motor de búsqueda Orión para su despliegue.

1.5.3. Visual Paradigm

Visual Paradigm es una herramienta CASE[14]: Ingeniería de Software Asistida por Computación. La misma propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación. Visual Paradigm ha sido concebida para soportar el ciclo de vida completo del proceso de desarrollo del software a través de la representación de todo tipo de diagramas. ("Visual Paradigm Essential" 2016)

Visual Paradigm posee varias características que lo hacen ideal para el desarrollo de los diagramas correspondientes. Como es la disponibilidad en múltiples plataformas. Posee un diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad. Por último y muy importante es de licencia gratuita y comercial.

1.5.4. Jmeter

Aplicación de escritorio Java desde la Fundación Apache *Software* diseñado para cargar la conducta funcional de prueba y medida de rendimiento. Originalmente diseñado para aplicaciones *Web* pruebas, pero desde entonces se ha expandido a otras funciones de prueba; puede ser utilizado para probar el rendimiento tanto en los recursos estáticos y dinámicos (archivos, *servlets*, *scripts* de Perl, objetos Java, bases de datos y consultas, servidores FTP y más). Puede ser utilizado para simular una carga pesada en el servidor, red u objeto poner a prueba su fuerza o para analizar el rendimiento general bajo diferentes tipos de carga; puede hacer un análisis gráfico de rendimiento o prueba de comportamiento del servidor / *script* / objeto bajo carga pesada concurrente.

1.5.5. Acunetix

Acunetix es un escáner de vulnerabilidades de aplicaciones web. La herramienta está diseñada para encontrar agujeros de seguridad en las aplicaciones web de la organización que un atacante podría aprovechar para obtener acceso a los sistemas y datos. Con acunetix se puede hacer *sniffing*, y nos debe de servir para poder proteger nuestros sitios de ataques de *hackers* que tienen las páginas *web* a su disposición las 24 horas del día para escanearlas y afectarnos en las vulnerabilidades que tengamos.

[14] Son diversas Aplicaciones informáticas destinadas a aumentar la productividad en el Desarrollo de software. Ingeniería de Software Asistida por Computadoras

1.6. Metodología de desarrollo.

Una metodología de desarrollo de software se refiere al entorno que se usa para estructurar, planificar y controlar el proceso de desarrollo de un sistema de información. Una metodología de desarrollo de software consiste en una filosofía con una base de procesos. La metodología consiste en múltiples herramientas, modelos y métodos, para asistir en el proceso de desarrollo (SlideShare 03:04:28 UTC).

Para el desarrollo de la solución se decide utilizar la metodología OpenUp. De tipo ágil, está dirigida a la gestión y desarrollo de proyectos de software basados en desarrollo iterativo, ágil e incremental. Apropiada para proyectos pequeños y de bajos recursos. Posee varias iteraciones durante el ciclo de vida del proyecto, que no superan las pocas semanas de duración, en dependencia a los acuerdos que se tomen en el equipo de trabajo. Se debe tener en cuenta que cada iteración concluye obligatoriamente con una muestra concreta del producto, que necesariamente tiene que ser “demostrativa” o “explotable”, ya que es la forma que tiene la metodología de desarrollo de demostrarle el valor agregado al cliente (Cabrera González and Pompa Torres 2012). Además, es la metodología que guía el proceso de desarrollo de software en el proyecto Orión.

1.7. Conclusiones del capítulo.

Partiendo de los elementos teóricos expuestos en este capítulo, se puede llegar a las siguientes conclusiones.

1. El estudio de los conceptos asociados a la recuperación de información y las noticias digitales en la Web dan como resultado un mayor entendimiento de los componentes de la investigación.
2. Al analizar los sistemas de búsqueda de noticias más utilizados en el mundo, se identificaron requerimientos necesarios para implementar la propuesta de solución.
3. Debido a las consecuencias que trae la falta de un sistema de búsqueda de noticias especializado en la red cubana, como sería el gasto que supone el uso de buscadores externos, el limitado acceso a sistemas internacionales y los resultados que muestran, se hace necesario desarrollar un sistema que permita la recuperación de noticias publicadas en la web cubana.

Capítulo 2. Diseño del Subsistema de búsqueda de noticias del buscador Orión

Un flujo de trabajo es un conjunto de actividades relacionadas que producen resultados visibles que pueden ser analizados. En este capítulo se realiza una descripción del sistema a desarrollar, a través de distintos diagramas donde se evidencian los distintos flujos que presenta la solución. Se abordarán los aspectos fundamentales para el diseño del sistema a implementar tomando como guía y haciendo uso de los artefactos que propone la metodología OpenUp. Además, se representa el modelo de dominio, así como los principales procesos mediante casos de uso, diagramas de clases del diseño, de colaboración y despliegue.

2.2. Modelo de dominio

Un modelo del dominio es una representación de las clases conceptuales del mundo real, no de componentes de software. No se trata de un conjunto de diagramas que describen clases de software, u objetos de software con responsabilidades, sino que modela clases conceptuales significativas en un determinado problema (Craig 2003). Teniendo en cuenta que la definición de procesos y roles del negocio llega a ser difícil, se hace necesario describir el funcionamiento de la aplicación mediante una serie de conceptos, entidades y sus relaciones, agrupándose en un modelo de dominio con el fin de contribuir a la comprensión del contexto del sistema.

En el diagrama modelo de dominio (ilustración 3) se inicia el proceso de búsqueda y recuperación partiendo del rastreo de los contenidos web, de los cuales en caso de ser noticias se envían sus datos al indexador para almacenarlos. El usuario puede introducir algún criterio de búsqueda en la aplicación web, la cual se encarga de comparar con los datos de los índices de cada indexador en busca de coincidencias, de encontrarse se mostrarían al usuario los resultados correspondientes.

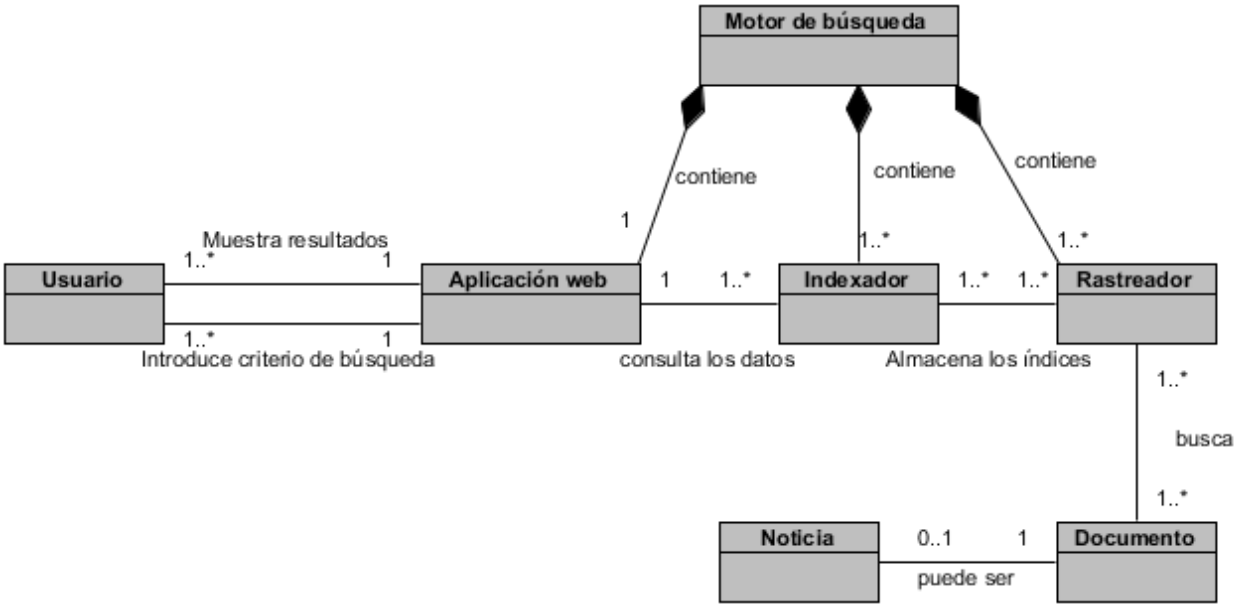


Ilustración 3: Diagrama Modelo de dominio

Tabla 2: Descripción de clases del modelo de dominio

Clases	Descripción
Usuario	Es la persona que realiza una búsqueda sobre un tema que le interese, a través de una interfaz web, al introducir un criterio de búsqueda.
Aplicación web	Es la parte del buscador encargada de interactuar con el usuario; recupera datos introducidos para encontrar coincidencias del criterio de búsqueda con los datos almacenados en el Índice. Además, de mostrarle los resultados al usuario.
Indexador	Es la parte del motor de búsqueda encargada de registrar los datos resultantes de la búsqueda del rastreador en forma de índice. También es el sistema con el cuál actúa directamente la aplicación web ya que es donde se buscarán coincidencias para el criterio y se devolverá como resultado todos los documentos o noticias que contengan el criterio solicitado.
Rastreador	Es la parte del buscador encargado de rastrear las noticias en la web y enviar al indexador la información.
Motor de búsqueda	Es el Sistema de recuperación de información en el que se aglomeran todos los componentes conectados entre sí, los cuales son: una aplicación web, uno o varios rastreadores e indexadores.
Noticia	Tipo de contenido del cual se almacenarán sus datos tal como su título, imagen, fuente, categoría, fecha de publicación, autor y descripción para mostrarlos al usuario.
Documento	Contenido electrónico encontrado en los servidores públicos que pueden o no ser noticias.

2.3. Descripción del sistema propuesto

El subsistema a desarrollar tiene como objetivo facilitar las búsquedas de noticias publicadas en la red cubana, realizadas a través del motor de búsqueda Orión. El sistema debe ser capaz de permitir a los usuarios realizar búsquedas usando filtros como los descritos a continuación:

Fecha de publicación: El sistema debe ser capaz de permitir la búsqueda de la noticia por una fecha determinada o un rango válido de publicación.

Medio o fuente de la Noticia: A través de este criterio se debe permitir mostrar las noticias de una fuente específica (**Ej**. Solo mostrar noticias provenientes de Cubadebate).

Temática o tipo de noticia: En este caso el sistema debe permitir el filtrado según el tipo de noticia que se desee obtener (deportes, salud, militares, cultura, políticas, internacionales, policiales).

Autor: Mediante este filtro el sistema debe permitir que se muestren las publicaciones realizadas por un determinado autor.

Idioma: El sistema debe mostrar las noticias realizadas en el idioma seleccionado, español o inglés.

Como resultado se creará un subsistema que permitirá realizar dos tipos de búsqueda de noticias. La primera de forma básica que se podrá hacer introduciendo un criterio de búsqueda en un campo de texto, criterio a partir del cual se devolverá todas las noticias que correspondan con él. El otro tipo de búsqueda es avanzado el cual permitirá utilizar los distintos filtros descritos anteriormente y devolverá resultados acordes a lo solicitado. Además, se podrán observar las noticias de manera automática al entrar al subsistema.

2.4. Especificación de los requisitos de software

Los requisitos son características que debe cumplir el software para solucionar un problema de la vida real, permitiendo esclarecer lo que el sistema debe hacer, sus características fundamentales y restricciones. De manera general, estos requisitos expresan las necesidades objetivas que presentan los usuarios, ante un sistema que resuelve un problema en particular de un determinado dominio (Sommerville, 2005).

2.4.1. Requisitos funcionales

La especificación de requisitos funcionales (RF) documenta las operaciones y actividades que un sistema debe ser capaz de realizar. Estos deben incluir las descripciones de los datos que se introducen en el sistema, descripciones de las operaciones realizadas por cada vista de la aplicación, descripciones de flujos de trabajo realizadas por el sistema, descripciones de los informes del sistema u otras salidas, además de cómo el sistema cumple con los requisitos reglamentarios aplicables (Ofni Systems, 2014).

Tabla 3: Requisitos funcionales

Número	Requisito funcional	Prioridad
RF1	Realizar búsqueda simple de noticias	Alta
RF2	Filtrar noticias dado uno o varios criterios.	Alta
RF3	Mostrar fuente de la noticia	Baja
RF4	Mostrar fecha de la noticia	Baja
RF5	Mostrar autor de la noticia	Baja
RF6	Mostrar imagen asociada a la noticia (primera imagen)	Baja
RF7	Mostrar resumen de la noticia	Baja
RF8	Mostrar título de la noticia	Baja
RF9	Organizar por fecha de publicación	Baja
RF10	Identificar título de la noticia	Media
RF11	Identificar fecha de la noticia	Media
RF12	Identificar autor de la noticia	Alta
RF13	Identificar idioma de la noticia	Alta
RF14	Identificar fuente de la noticia	Media
RF15	Identificar ámbito de la noticia	Alta
RF16	Identificar temática de la noticia	Alta
RF17	Almacenar datos de noticia	Media

2.4.2. Requisitos no funcionales

Los requisitos no funcionales (RNF) son restricciones de los servicios o funciones ofrecidas por el sistema. Incluyen restricciones de tiempo, sobre el proceso de desarrollo y estándares. Los requerimientos no funcionales a menudo se aplican al sistema en su totalidad. Apenas se aplican a características o servicios individuales del sistema. Los requerimientos no funcionales, son aquellos que no se refieren directamente a las

funciones específicas que proporciona el sistema, sino a las propiedades emergentes de éste como la fiabilidad, el tiempo de respuesta y la capacidad de almacenamiento. De forma alternativa, definen las restricciones del sistema como la capacidad de los dispositivos de entrada/salida y las representaciones de datos que se utilizan en las interfaces del sistema.

Tabla 4: Requisitos no funcionales

Categoría	Número	Descripción
Software	RnF1	El sistema requiere de un sistema operativo Unix.
	RnF2	El sistema requiere la instalación del servidor web y de Servlet Tomcat 7 para el correcto funcionamiento del servidor de Solr.
	RnF3	El sistema requiere la instalación de la Máquina Virtual de Java (JVM, por sus siglas en inglés) para el correcto funcionamiento del rastreador.
	RnF4	El sistema requiere la instalación del servidor web Apache en su versión 2.4 y PHP 5.5 o superior para poder visualizar la interfaz web.
Hardware	RnF5	El servidor de índice requiere mínimo: 4GB RAM, CPU Core i3 250 GB HDD.
	RnF6	El servidor de rastreo requiere mínimo: 4GB RAM, CPU Core i3 250 GB HDD.
	RnF7	El servidor de la aplicación se requiere mínimo: 4GB RAM, CPU Core i3 80 GB HDD.
Rendimiento	RnF13	El sistema debe poseer un buen rendimiento, lo cual se encuentra estrechamente vinculado a los requerimientos de *hardware* y *software* del sistema, los cuales fueron detallados. Debe tener un tiempo de respuesta entre 0 y 4 segundos.

2.5. Modelo de casos de uso del sistema.

El diagrama de casos de uso representa la forma en cómo un Actor[15] opera con el sistema, además de la forma y orden en como los elementos interactúan. Es un diagrama que muestra la relación entre los actores y los casos de uso del sistema (CU[16]). Los RF del sistema, definidos anteriormente quedan agrupados en 4 CU organizados en la tabla 5.

Tabla 5: CU del sistema

Referencia a requisitos	Nombre del caso de uso

[15] El actor representa una entidad externa que interactúa con el sistema.
[16] Técnica para la captura de requisitos potenciales de un nuevo sistema.

RF1	CU 1: Buscar noticia de forma simple.
RF3, RF4, RF5, RF6, RF7, RF8, RF9	CU 2: Visualizar noticias.
RF2	CU 3: Buscar noticia de forma avanzada.
RF10, RF11, RF12, RF13, RF14, RF15, RF16	CU 4: Recuperar[17] noticias.

2.5.1. Caso de uso

En ingeniería del software, un caso de uso (CU) es una técnica para la captura de requisitos potenciales de un nuevo sistema o una actualización de software. Los mismos proporcionan uno o más escenarios que indican cómo debería interactuar el sistema con el usuario o con otro sistema para conseguir un objetivo específico, además son una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal. El caso de uso hace referencia al sistema a construir, detallando su comportamiento, el cual se traduce en resultados que pueden ser observados por el actor. Describen las acciones o tareas que los actores quieren que el sistema haga, por lo que un caso de uso debería ser una tarea completa desde la perspectiva del actor. Seguidamente se presenta el modelo de casos de uso que permitirá entender las funcionalidades del sistema propuesto.

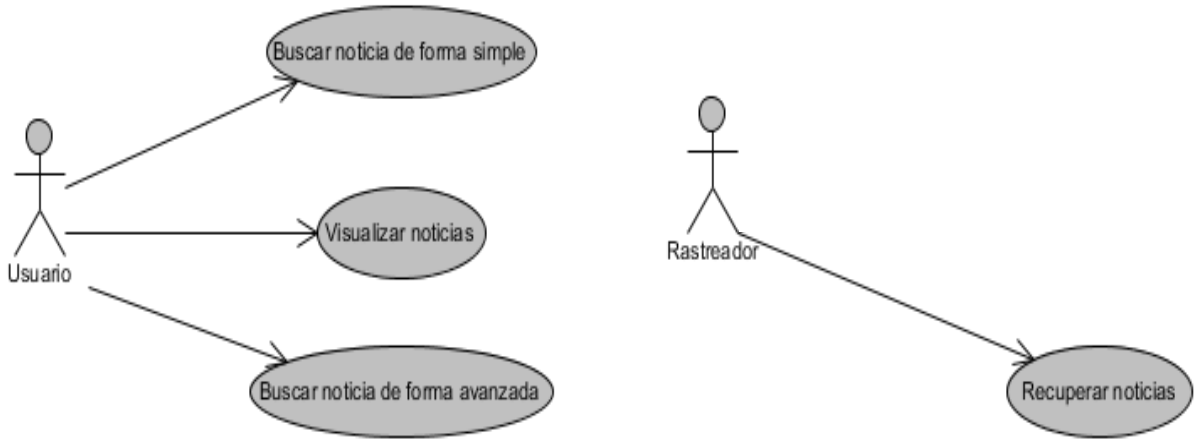


Ilustración 4: Diagrama de casos de uso

[17] Recuperar noticias (Obtener los datos de las noticias)

2.5.2. Especificación de casos de uso

La especificación de los casos de uso se refiere a la descripción de cada una de las partes definidas en el diagrama de casos de uso para lograr su descripción completa. El comportamiento de un CU se puede especificar describiendo un flujo de eventos de forma textual, lo suficientemente claro que lo entienda fácilmente el programador. Es importante especificar cómo y cuándo se empieza y acaba el caso de uso (Rodríguez, 2014).

Tabla 5: CU Recuperar noticias

Objetivo	Con este CU el rastreador debe ser capaz de identificar y de indexar los datos de la noticia.	
Actores	Rastreador.	
Resumen	El rastreador extrae y procesa los datos de cada noticia encontrada en la Web para enviárselos al indexador el cual crea el índice.	
Complejidad	Alta.	
Prioridad	Alta.	
Precondiciones	Una persona responsabilizada por el cliente, debe haber configurado e iniciado el mecanismo de rastreo.	
Postcondiciones	Se obtienen todos los datos de las noticias y se almacenan.	
Flujo de eventos		
Flujo básico Recuperar noticias		
	Actor	**Sistema**
1.	Inicia el mecanismo de rastreo.	
2.		Recupera el contenido de los documentos.
3.		Envía los datos al indexador.
4.		Se crea el índice en el indexador. **Finaliza el CU.**
Relaciones	CU Incluidos	No procede

	CU Extendidos	No procede

2.6. Diagrama de clases del diseño

El diagrama de clases del diseño presenta las clases del sistema con sus relaciones estructurales y de herencia. En el caso de las aplicaciones Web, el diagrama de clases representa las colaboraciones que ocurren entre las páginas, donde cada página lógica puede ser representada como una clase. Este describe gráficamente las especificaciones de las clases de software y las interfaces en una aplicación. (“Diagramas de Clases” 2011).

En la siguiente ilustración se muestran las clases que intervienen en el proceso de analizar una noticia. Para ello, el rastreador se comunica a través de la interfaz Parse con la clase NoticiaParser, la cual es la encargada de devolver los metadatos extraídos de la noticia. Esta clase depende de NoticiaMetadataExtractor para la extracción de los metadatos; los cuales, en un subproceso independiente son procesados por NoticiaIndexingFilter para decidir cuáles metadatos y cómo serán indexados. Esta última clase se comunica con el rastreador mediante la interfaz IndexingFilter.

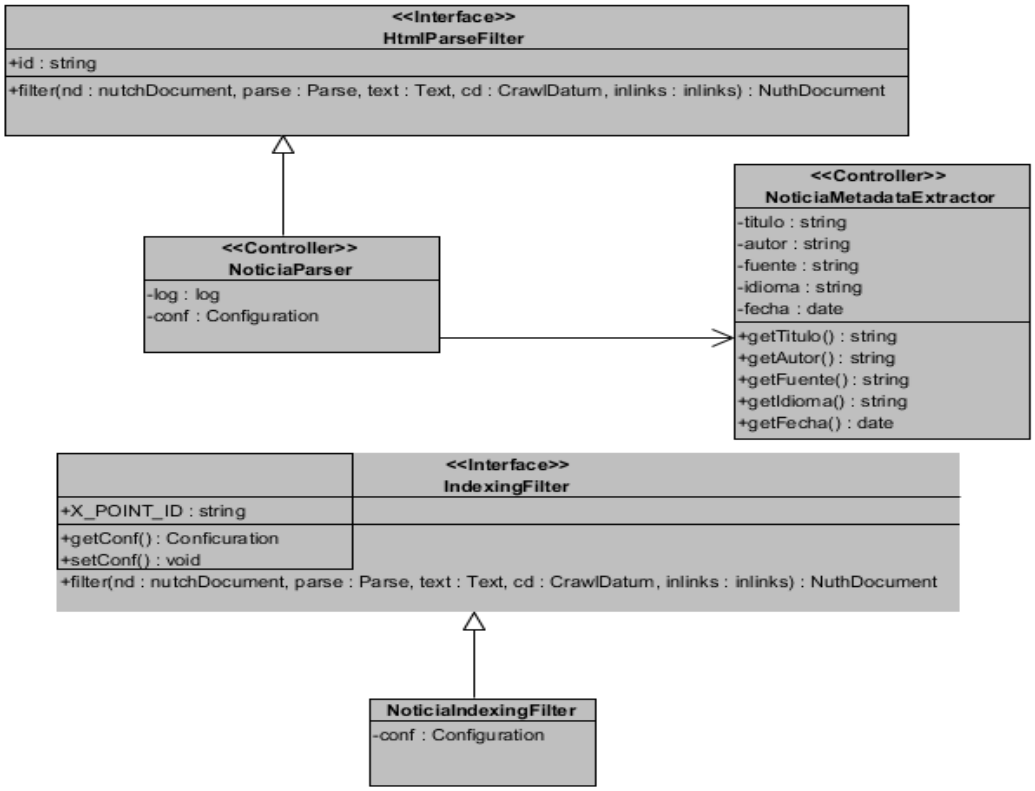


Ilustración 5: Diagrama de clases del diseño del CU Capturar datos de la noticia

2.7. Diagramas de colaboración

Un diagrama de colaboración explica gráficamente las interacciones existentes entre las instancias (y las clases). Describen las interacciones entre los objetos en un formato de grafo o red. (Larman, 2004). El siguiente diagrama de colaboración corresponde al CU Recuperar noticias. El mismo es iniciado por el rastreador, se envía una instancia de ParserResult con los datos de los documentos encontrados en la Web. El objeto FeedIndexingFilter obtiene estos datos de ParserResult y a su vez envía estos datos obtenidos a la clase NucthDocument, la cual representa el documento a indexar.

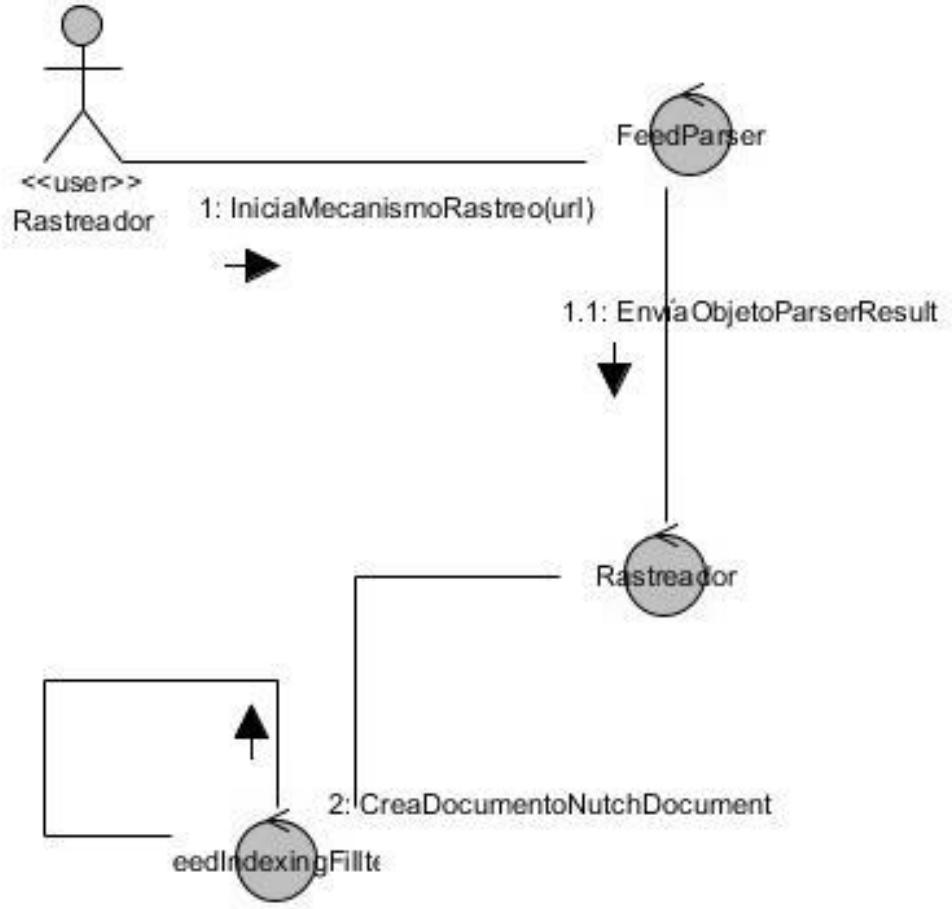


Ilustración 6: Diagrama de colaboración CU Recuperar noticias

2.8. Arquitectura del sistema propuesto

El funcionamiento interno de Symfony 2.7 está basado en la arquitectura Modelo-Vista-Controlador (MVC). Sin embargo, este solo proporciona herramientas para el controlador y la vista, dejando el modelo por parte del usuario (Potencier, 2011). A continuación, se muestra la arquitectura del sistema.

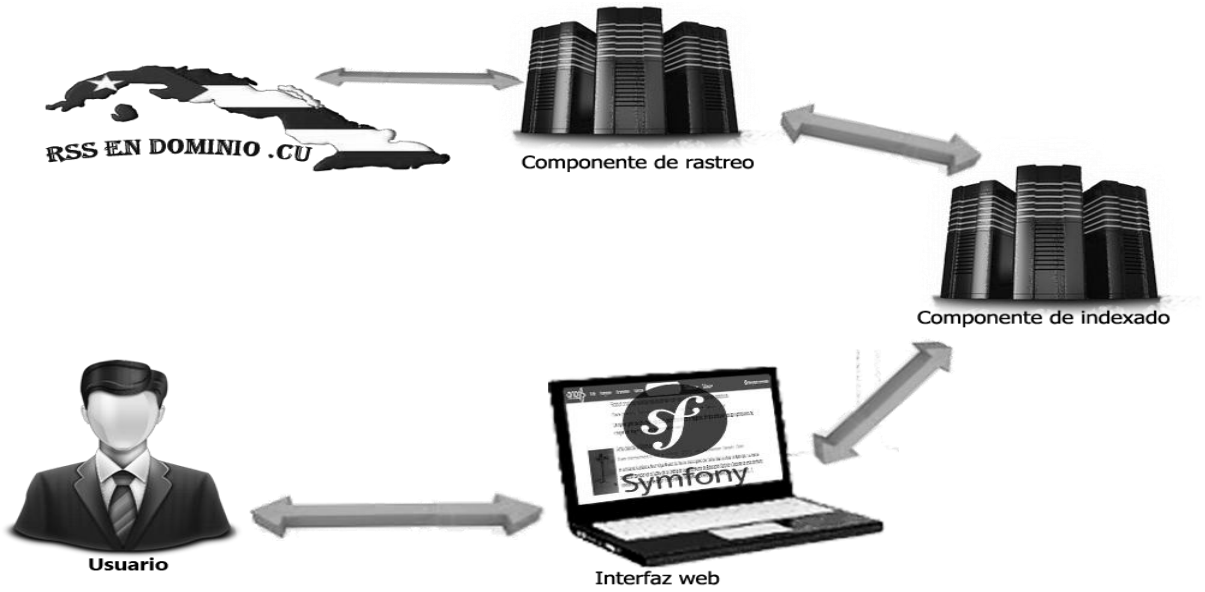


Ilustración 7: Arquitectura del sistema propuesto

Como se observa en la ilustración anterior, el sistema propuesto estará compuesto por una aplicación web encargada de atender las peticiones de los usuarios, consultar los datos almacenados en el indexador y mostrar los resultados. La aplicación es la encargada de interactuar con el usuario, así como mostrarle las noticias obtenidas a partir del indexador, encargado de almacenar en un índice creado por él, los datos encontrados en la Web por el rastreador que correspondan al criterio de búsqueda.

2.8.1. Patrones utilizados en el desarrollo de software

Los patrones de diseño representan la descripción de un problema particular y recurrente, que aparece en contextos específicos, y presenta un esquema genérico demostrado con éxito para su solución; este último se especifica mediante la descripción de los componentes que la constituyen, sus responsabilidades y desarrollos, así como también la forma como estos colaboran entre sí (Larman, 2004).

En el diseño del subsistema de búsqueda de noticias se tuvieron en cuenta los siguientes patrones GRASP (Patrones Generales de Software para Asignación de Responsabilidades), que describen los principios fundamentales de la asignación de responsabilidades a objetos:

Experto: Este patrón plantea que se debe asignar una responsabilidad al experto en información, en otras palabras, a la clase que cuenta con los datos necesarios para cumplir la responsabilidad. De esta forma, se conserva el encapsulamiento de la información, puesto que los objetos ejecutan las tareas que le corresponden de acuerdo a la información que poseen, esto conlleva a un bajo acoplamiento, lo que da lugar a sistemas más robustos y fáciles de mantener (Larman, 2004). Siguiendo el patrón Experto se le asignaron responsabilidades determinadas solamente a las clases que cuentan con la información necesaria para dar cumplimiento a las mismas. El mismo se pone de manifiesto como a continuación se refleja:

Tabla 6: Asignación de responsabilidades

Clase objeto	**Responsabilidad**
FeedParser	Esta clase tiene las herramientas necesarias para extraer la información de las noticias.

Creador: La instanciación de una clase es una de las actividades fundamentales en un sistema orientado a objetos. Este patrón guía la asignación de responsabilidades relacionadas con la creación de objetos, con lo que se logra menos dependencia y mayores oportunidades de reutilización de código (Larman, 2004). Ejemplo de esto es la clase controladora FeedParser que crea un objeto de la clase SAXBuilder para construir el archivo .xml (Ver ilustración 9).

```
//Read file categories.xml and convert categories into one
List<List<String>> categorias = new ArrayList<List<String>>();
SAXBuilder builder = new SAXBuilder();
File xmlFile = new File("/opt/nutch_amaury/runtime/local/conf/categories.xml");

try {
    Document document = (Document) builder.build(xmlFile);
```

Ilustración 9: Ejemplo Clase FeedParser. Patrón Creador

2.9. Modelo de datos

El modelado de datos se basa en la identificación de los objetos primarios que va a procesar el sistema, la composición y atributos de los mismos. Además, de dónde se encuentran almacenados actualmente dichos objetos, la relación entre ellos y los procesos que los transforman. A continuación se muestra el modelo de datos para el sistema que se propone desarrollar.

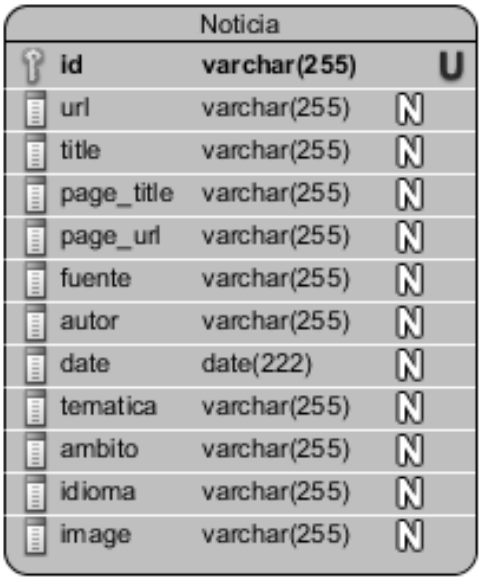


Ilustración 10: Modelo de Datos

Descripción de los campos del modelo de datos

1. **id**: identificador del documento o noticia almacenada, que corresponderá con la URL por ser única.
2. **url**: ubicación relativa en la web donde se encuentra el documento correspondiente esta es de tipo única ya que un url corresponde a un único recurso.
3. **título**: corresponde al título de la noticia.
4. **page_title**: corresponde al título de la página donde se encuentra la noticia.
5. **page_url**. Corresponde a la url del servidor de la página.
6. **fuente**: nombre del periódico o sitio donde se encuentra la noticia.
7. **autor**: nombre del creador de la noticia.
8. **date:** corresponde a la fecha de publicación de la noticia.
9. **temática:** corresponde a la clasificación o categoría de la noticia.
10. **idioma:** corresponde al idioma de publicación.

11. **imagen:** corresponde a la primera imagen de contenida en la noticia.

2.10. Diagrama de despliegue

El diagrama de despliegue es utilizado para capturar los elementos de configuración del procesamiento y las conexiones entre dichos elementos. Es utilizado también para visualizar la distribución de los componentes de software en los nodos físicos. La relación entre los nodos es denominada como protocolos de comunicación (SparxSystems, 2014).

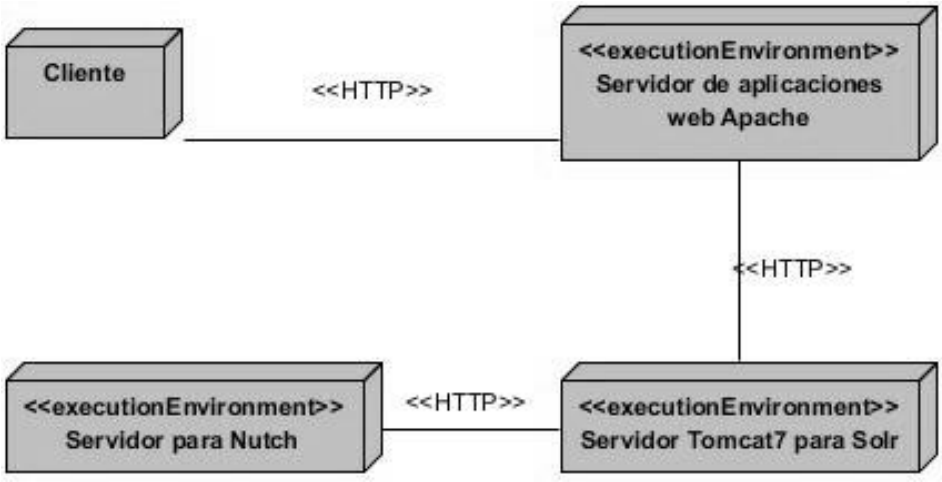


Ilustración 11: Diagrama de despliegue

Como se puede apreciar en la imagen anterior el cliente representa los usuarios que desde distintos dispositivos podrán realizar búsquedas de noticias, a través del protocolo HTTP, haciendo uso de un navegador web. El nodo "Servidor de aplicaciones web Apache" es el encargado de atender y dar respuesta a cada una de las solicitudes del cliente. Además, se observan dos nodos más, uno que representa el servidor maestro para Solr con Tomcat7 como contenedor de *servlets*[18] y otro donde deberá ser instalado el servidor maestro Nutch. Se sugiere que se encuentren en servidores independientes con el objetivo de utilizar al máximo las características de hardware y software de estos, aunque pudieran alojarse en uno solo.

2.11. Conclusiones parciales

En este capítulo se abordaron una serie de aspectos correspondientes al análisis y diseño de la herramienta de búsqueda de noticias publicadas en la web llegándose a las siguientes conclusiones:

[18] Clase en el lenguaje de programación Java, utilizada para ampliar las capacidades de un servidor.

1. La representación y descripción de los artefactos generados garantizaron un mejor entendimiento de los flujos de trabajos presentes en el proceso de búsqueda de una noticia.
2. La identificación de los requisitos funcionales y no funcionales permitió definir las características y las condiciones que debe suministrar el subsistema.
3. La definición de la arquitectura y los patrones de diseño a utilizar, permitieron establecer las bases para fomentar la reutilización y las buenas prácticas de programación entre los desarrolladores durante la fase de implementación, así como disminuir el impacto de los cambios futuros en el código fuente.
4. La elaboración del diagrama de despliegue permitió identificar la disposición física de los artefactos de la herramienta informática a desarrollarse.

Capítulo 3. Implementación y pruebas del subsistema de búsqueda de noticias para el buscador Orión

La implementación es una fase importante en el desarrollo de software, puesto que esta representa la materialización o cumplimiento de todos los artefactos diseñados y descritos en el proceso de análisis y diseño de la aplicación, en forma de código. Dicho de otra forma, es la creación de la solución propuesta.

Seguidamente a este proceso, la solución debe ser sometida a diferentes pruebas con el fin de validar la calidad de esta. Esto se hace al crear distintas pruebas para comprobar que se cumple con todos los requisitos propuestos en las anteriores fases. Por consiguiente, en este capítulo se diseñarán y realizarán las pruebas necesarias para demostrar el cumplimiento de los requisitos tanto funcionales como no funcionales.

3.1. Diagrama de componentes

Un diagrama de componentes muestra las dependencias entre los ellos. Los componentes físicos incluyen archivos, bibliotecas compartidas, módulos, ejecutables, o paquetes. Describe también cómo se organizan los componentes de acuerdo con los mecanismos de estructuración y modularización disponible en el entorno de implementación y en el lenguaje de programación utilizado. (Pressman 2005)

A continuación, se muestra el diagrama de componentes del subsistema de búsqueda de noticias del buscador cubano Orión.

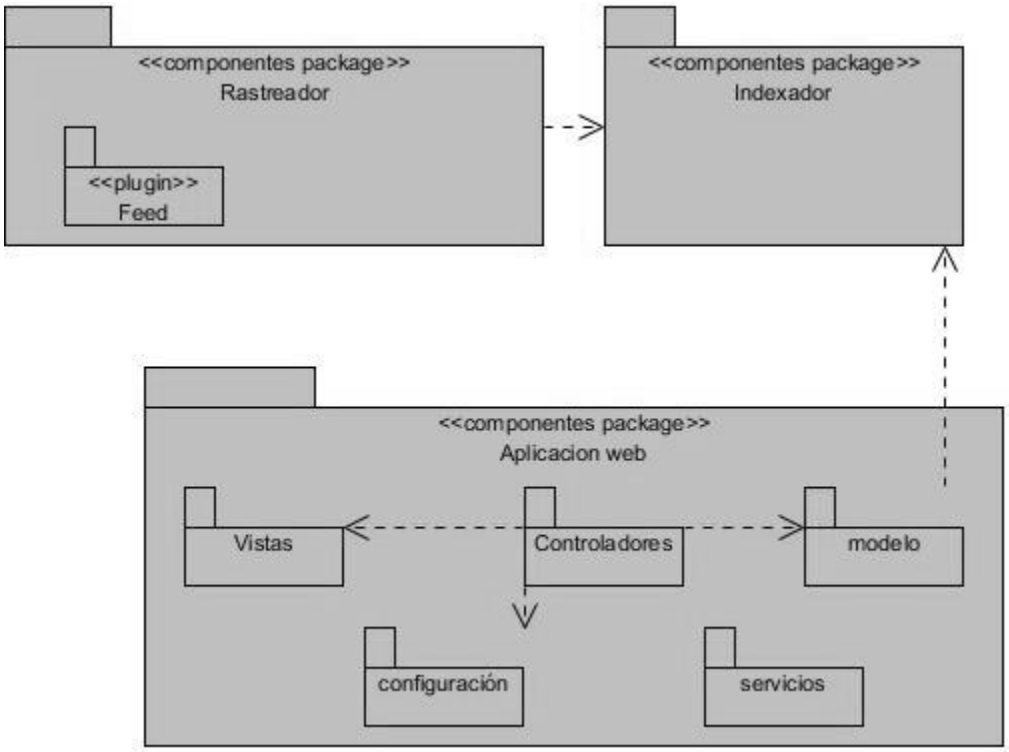


Ilustración 12: Diagrama de componentes del sistema

El diagrama anterior consta de tres paquetes que representan los componentes del subsistema de búsqueda de noticias del buscador Orión. Las responsabilidades de cada uno de estos son:

Rastreador: Encargado de implementar el proceso de recuperación de noticias en la web.
Indexador: Contiene las configuraciones necesarias para el indexado de los datos y establece la comunicación con el rastreador y la aplicación web.
Aplicación web: Contiene los paquetes controladores, vistas, servicios, formularios y configuración.

El paquete Rastreador está compuesto por dos paquetes, Parsenoticia y Recoverynews; el primero es el *plugin* utilizado por Nutch para la recuperación a partir de canales RSS, al cual se le realizaron cambios con el objetivo de obtener más información, el segundo corresponde al desarrollado para recuperar las noticias directamente.

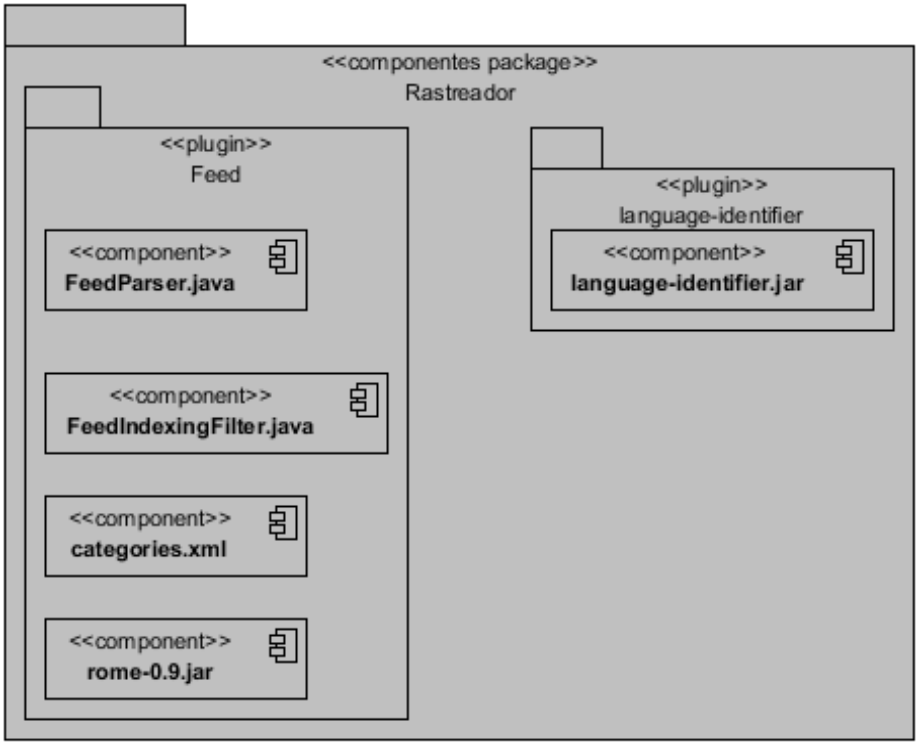


Ilustración 13: Diagrama de componentes del paquete Rastreador

El paquete Feed está compuesto por dos componentes fundamentales:

FeedParser.java: Representa al fichero que contiene la clase FeedParser encargada de extraer los datos de las noticias publicadas en canales RSS, enviándolos al rastreador.
FeedIndexingFilter.java: Componente con la responsabilidad de decidir qué datos y de qué manera se indexarán.
Categories.xml: es un archivo que cuenta con una estructura de categorías padres de subcategorías con el objetivo de converger en una sola categoría estándar.
Rome-0.9.jar: es la librería encargada de conectar con los canales RSS para obtener sus datos.
language-identifier.jar: pertenece al plugin del mismo nombre, es el encargado de identificar el lenguaje de las noticias.

Los componentes que forman el paquete Indexador son ficheros de configuración que garantizan el correcto almacenamiento de los datos, así como la comunicación con el resto de los componentes de la arquitectura del sistema.

Ilustración 14: Diagrama de componentes del Indexador

schema.xml: Contiene los campos o atributos de un documento, así como los filtros que se le aplican a dichos campos, los cuales ayudan a mejorar y facilitar la indexación y la búsqueda.
solrconfig.xml: Fichero de configuración principal de Solr.

Como se muestra en la siguiente imagen, los componentes físicos del paquete Aplicación web están divididos en seis subpaquetes: Vistas, Controladores, Formularios, Servicios, Modelo y Configuración

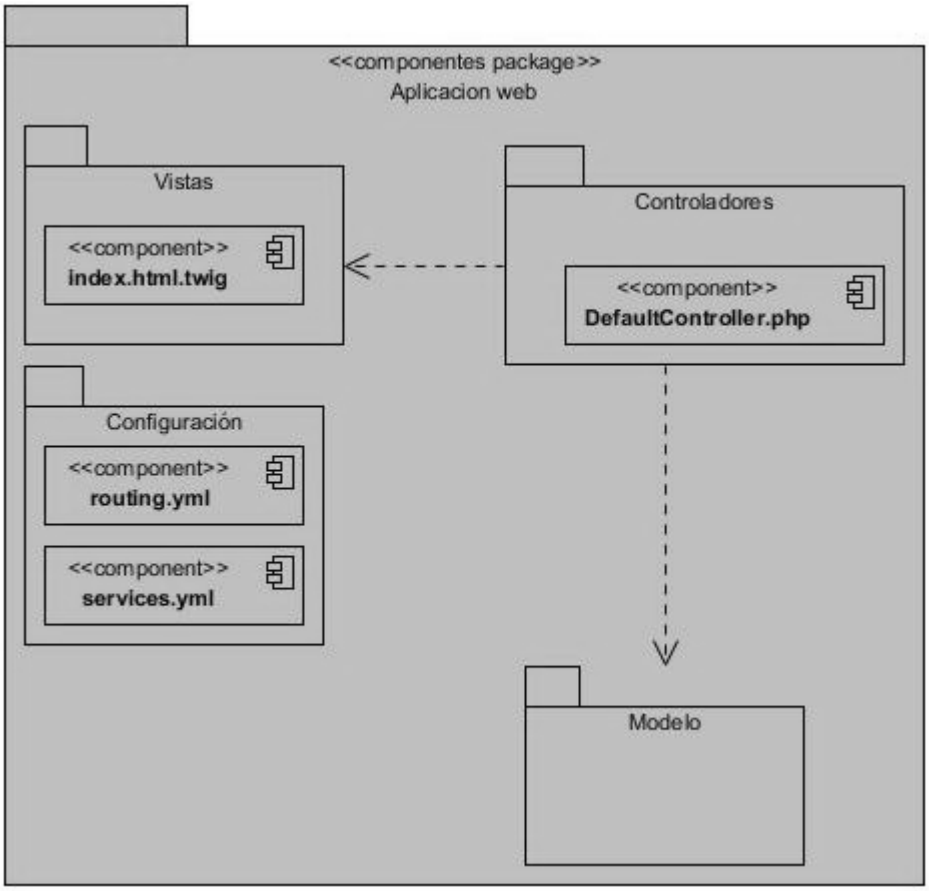


Ilustración 15: Diagrama de componentes de la aplicación web

En el paquete Controladores se incluye el componente DefaultController.php, el cual contiene la clase newsController, encargada de controlar las peticiones de los usuarios, crear las vistas correspondientes a cada una de ellas y manejar los servicios del sistema, así como los formularios de búsqueda.

Modelo, es el paquete encargado de la comunicación con el componente de indexación y la representación de la información almacenada en él.

En el paquete Configuración se encuentran los ficheros de configuración correspondientes a la aplicación web:
routing.yml: Contiene las rutas que utilizará el sistema
services.yml: Contiene la configuración de los servicios del sistema.
El paquete Vistas incluye las vistas que muestra el sistema al usuario, las cuales son:
index.html.twig: Es la vista principal del sistema, la cual muestra los formularios de búsqueda simple y búsqueda avanzada. Además de los resultados de las búsquedas.

3.2. Estándares de codificación

Un estándar de codificación son reglas que se siguen para la escritura del código fuente, de tal manera que a otros programadores se les facilite entender el código. (Junta de Andalucía. Convenio de codificación específico para Drupal, 2013) Para facilitar el entendimiento del código y fijar un modelo a seguir, se establecieron los estándares de codificación, para la aplicación web el estándar PSR-2[19], y para los *plugins* de *nutch* los estándares definidos por la comunidad de java[20].

Organización de los ficheros

Un fichero consiste de secciones que deben estar separadas por líneas en blanco y comentarios opcionales que identifican cada sección. Serán evitados los ficheros de más de 2000 líneas pues son incómodos y en ocasiones provoca que la clase no encapsule un comportamiento claramente definido.

Identación

Se evitarán las líneas de más de 80 caracteres, ya que no son manejadas bien por muchas terminales y herramientas. Cuando una expresión no entre en una línea, se romperá de acuerdo a estos principios:

- Romper después de una coma.
- Romper antes de un operador.

Declaraciones

Se realizará una declaración por línea, ya que facilita los comentarios. Se inicializará las variables locales donde se declaran. Se pondrán las declaraciones solo al principio de los bloques para evitar confusión en programadores no preavisados. En las declaraciones de clases e interfaces se tendrá en cuenta no usar ningún espacio en blanco entre el nombre de un método y el paréntesis que abre su lista de parámetros. Los métodos se separarán con una línea en blanco.

Espacios en blanco

Las líneas en blanco mejoraran la facilidad de lectura separando secciones de código que están lógicamente relacionadas. Se hará uso de dos líneas en blanco entre secciones de un fichero fuente y entre las definiciones de clases e interfaces y una línea en blanco entre métodos, variables locales de un método y su primera sentencia y entre las distintas secciones lógicas de un método para facilitar la lectura.

[19] https://github.com/php-fig/fig-standards/blob/master/accepted/PSR-2-coding-style-guide.md
[20] http://systempix.com/descargas/Convenciones_Codigo_Java.pdf

3.3. Posicionamiento web y relevancia

La importancia de una página web es un asunto inherentemente subjetivo, que depende de los intereses de los lectores, sus conocimientos y actitudes. Pero todavía hay mucho que se puede decir de manera objetiva acerca de la importancia relativa de las páginas web.

Para medir la importancia relativa de las páginas web, se utiliza por la mayoría de los buscadores el PageRank, que es un método para calcular una clasificación para cada página web basado en la gráfica de la web. PageRank tiene aplicaciones en la búsqueda, exploración y la estimación de tráfico

PageRank o ranking de Page es un conjunto de algoritmos creados por Larry Page y Serguei Brin, usado por Google desde el 9 de enero de 1999. Los algoritmos son utilizados para cuantificar de forma numérica la relevancia de los documentos (páginas web) encontrados en Internet. PageRank otorga la medida de relevancia teniendo en cuenta la cantidad de enlaces que apuntan a una página, pero no lo mide solamente teniendo en cuenta la cantidad de enlaces encontrados, sino teniendo en cuenta la importancia de la página de la que parte el enlace, haciendo más importante el mismo; lo que significa que la página apuntada hereda de cierta manera parte de la importancia de quien la apunta.

El PageRank mide la importancia que posee la página en internet y es el indicador que utiliza Google para el posicionamiento, a mayor relevancia más arriba en el listado de resultados al realizar una búsqueda. Google ordena los resultados de la búsqueda utilizando el algoritmo PageRank. A cada página web se le asigna un número en función del número de enlaces de otras páginas que la apuntan, el valor de esas páginas y otros criterios no públicos.

El proyecto Orión utiliza el cálculo de la relevancia de los documentos que implementa Lucene en el servidor de índices Solr.

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} \left(\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot \text{t.getBoost}() \cdot \text{norm}(t,d) \right)$$

El *scoring* es el proceso de determinar cuánto de relevante es un documento con respecto a la consulta del usuario. El usuario realiza una búsqueda y para cada documento que contiene esas palabras se calcula un número al que se le suele llamar score, peso o relevancia, que se calcula según se describe mediante la

fórmula conceptual del scoring de Lucene. Los factores involucrados en el algoritmo TFIDF implementado en Lucene son:

tf: term frequency en un document, medida de cuán frecuente un término aparece en un documento.
idf: inverse document frequency, medida de cuán frecuente un término aparece en todo el índice.
coord: número de términos de la consulta que fueron encontrados en el documento.
boost (index|query) = factor de aumento (o disminución) para la ocurrencia de un término en un campo determinado.

Es necesario destacar que excepto el parámetro de *boost* que se ha personalizado, de forma tal de dar prioridad a documentos (url) que tengan los términos introducidos en la consulta determinados campos, los demás factores son los que provee Lucene/Solr por defecto calculados mediante el *scoring*.

3.4. Validación del sistema

Una vez terminada la implementación del producto que se requiere es necesario realizarle pruebas con el objetivo de detectar errores en la aplicación; este proceso resulta de gran importancia ya que da una medida de la calidad del mismo siempre que se lleve a cabo de la forma correcta.

3.4.1. Pruebas funcionales

Las pruebas funcionales, también denominadas pruebas de comportamiento se centran en los requisitos funcionales del software. Estas permiten al desarrollador obtener un conjunto de datos de entrada que evalúan todos los requisitos funcionales del sistema. Por esto se denominan pruebas funcionales, donde se suministran datos de entrada y se observa la salida, sin necesidad de conocer el funcionamiento interno del software.

Dentro de esta clasificación se encuentra la técnica partición equivalente el cual es un método de prueba de caja negra que divide el dominio de entrada de un programa en clases de datos de los que pueden derivarse casos de prueba, también (Gil, 2016):

- Se dirige a la definición de casos de prueba que descubran clases de errores, reduciendo así el número total de casos de prueba que hay que desarrollar.
- El diseño de casos de prueba para la partición equivalente se basa en una evaluación de las clases de equivalencia para una condición de entrada.

A continuación, se muestran algunos Casos de Pruebas Basados en requisitos que son de importancia para el cliente.

Tabla 7: Caso de prueba Filtrar por criterio de búsqueda

Escenario	Descripción	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Respuesta del sistema	Flujo central
EC 1.1 Filtrar por criterio	*En este escenario se realiza el filtrado de la noticia por criterio de búsqueda correctamente.*	*V*	*NA*	*NA*	*NA*	*NA*	*Muestra la noticia la cual contiene este criterio.*	*1. Seleccionar búsqueda avanzada.* *2. Introducir dato en campo de entrada criterio.* *3. Click en el botón buscar.*
		Septeto						
EC 1.1 Filtrar por criterio(carácter extraño)	*En este escenario no se realiza el filtrado de la noticia por criterio de búsqueda correctamente.*	*I*	*NA*	*NA*	*NA*	*NA*	*Muestra la página de error de Symfony.*	*1. Seleccionar búsqueda avanzada.* *2. Introducir dato en campo de entrada criterio.* *3. Click en el botón buscar.*
		%						

Tabla 8: Caso de prueba Filtrar por autor

Escenario	Descripción	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Respuesta del sistema	Flujo central
EC 1.1 Filtrar por autor	*En este escenario se realiza el filtrado de la noticia por autor correctamente.*	*NA*	*NA*	*V*	*NA*	*NA*	*Muestra la noticia la cual contiene este criterio.*	*1. Seleccionar búsqueda avanzada.* *2. Introducir dato en campo de entrada autor.* *3. Click en el botón buscar.*
				Redacción digital				
EC 1.1 Filtrar por autor(carácter extraño)	*En este escenario no se realiza el filtrado de la noticia por autor correctamente.*	*NA*	*NA*	*I*	*NA*	*NA*	*Muestra la página de error de Symfony*	*1. Seleccionar búsqueda avanzada.* *2. Introducir dato en campo de entrada autor.* *3. Click en el botón buscar.*
				%				

Tabla 9: Variables

No	Nombre de campo	Clasificación	Valor Nulo	Descripción

1	*Criterio*	*campo de texto*	*Si*	*En este campo el usuario introduce un criterio de búsqueda.*
2	*Fecha*	*datepicker*	*Si*	*Fecha o rango de fechas de la publicación de la noticia.*
3	*Autor*	*campo de texto*	*Si*	*En este campo el usuario introduce el creador de la noticia.*
4	*Fuente*	*select*	*Si*	*En este campo el usuario selecciona la posible fuente de la publicación.*
5	*Temática*	*select*	*Si*	*Se le permite al usuario seleccionar la categoría de la noticia.*

Resultados de las pruebas funcionales

Luego de ser revisados los casos de pruebas correspondientes a las funcionalidades del Subsistema de búsqueda de noticias del buscador Orión se detectaron un total de 10 no conformidades, las cuales fueron en su mayoría errores ortográficos y fueron debidamente resueltas. La siguiente gráfica muestra en dos iteraciones el resultado obtenido.

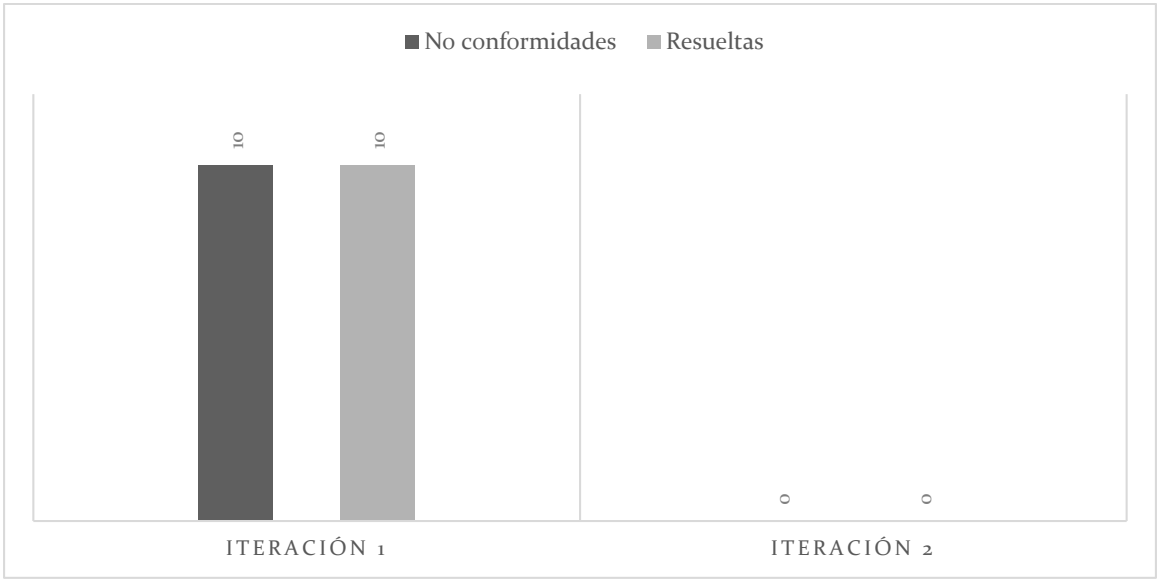


Ilustración 16: Resultados de las pruebas funcionales.

3.4.2. Prueba de integración

El proceso de integración del sistema implica construirlo a partir de sus componentes y probar el sistema resultante para encontrar problemas que pueden surgir debido a su integración. Una vez que los componentes han sido integrados, tiene lugar un extenso programa de pruebas del sistema. Estas pruebas pretenden probar las interfaces entre los componentes y el comportamiento del sistema en su totalidad (SOMMERVILLE 2005).

El motor de búsqueda cubano Orión, cuenta con una estructura que facilita la integración de nuevos subsistemas para incrementar sus funcionalidades. Por tal motivo, se decidió realizar pruebas de integración descendentes. La cual consiste en integrar el subsistema moviéndose hacia abajo por la jerarquía de control, comenzando por el programa principal. En la siguiente ilustración se identifica la estructura de implementación del buscador Orión y el subsistema NewsBundle integrado correctamente. Las pruebas de integración realizadas permitieron detectar problemas de conexión con el servidor de índices Solr (ver ilustraciones 18 y 19). Luego de instalar la librería Curl y solucionada esta no conformidad el subsistema desarrollado se integró de manera correcta al motor de búsqueda Orión (ver ilustración 20).

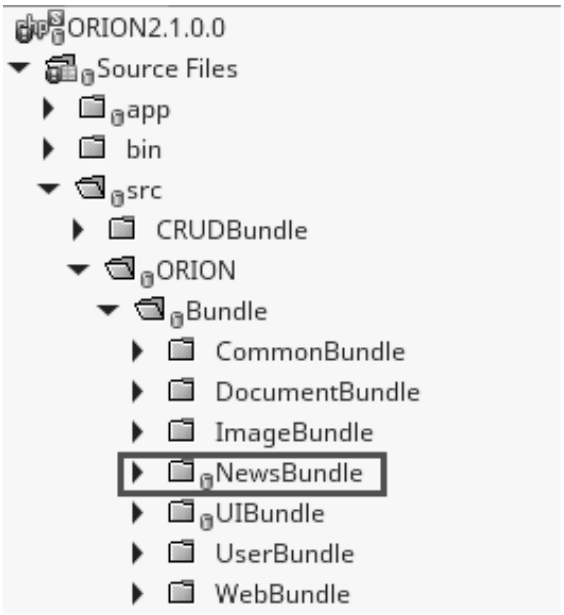


Ilustración 17: Estructura de implementación del sistema con el subsistema integrado.

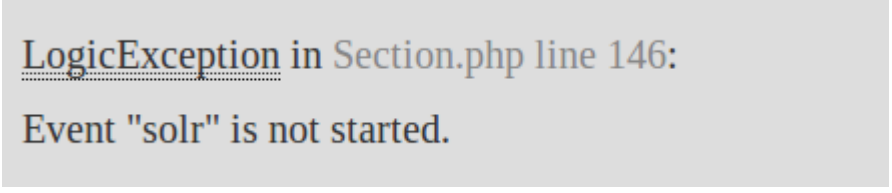


Ilustración 18: Error de Symfony (1).

HttpException in Curl.php line 195:

Solr HTTP error: HTTP request failed, Connection timed out after 5001 milliseconds

Ilustración 19: Error de Symfony (2).

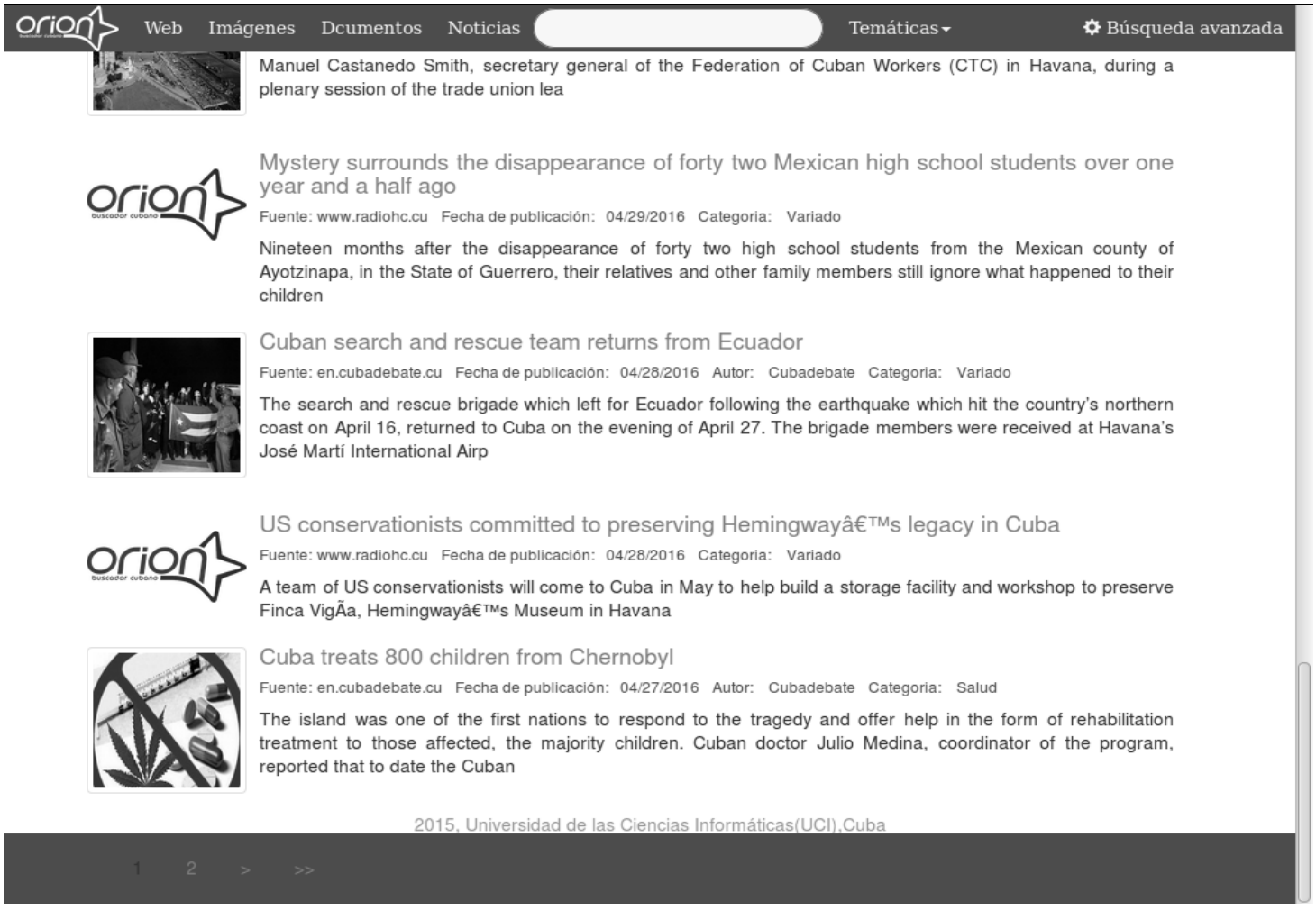


Ilustración 20: Estructura de implementación del sistema con el subsistema integrado.

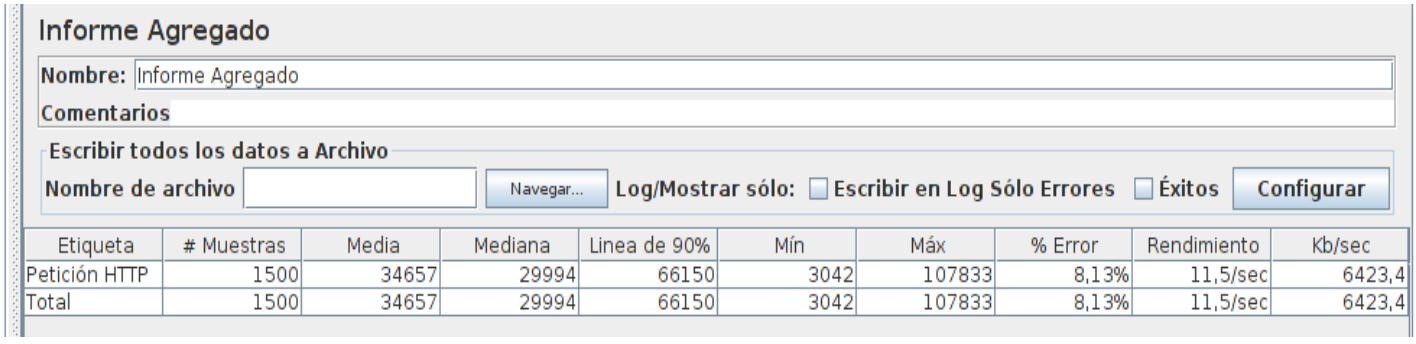
3.4.3. Pruebas de carga y estrés

Las pruebas de carga consisten en probar el funcionamiento del software bajo condiciones extremas. Estudia la especificación del software, las funciones que debe realizar, las entradas y las salidas, analizando los valores límites. Las pruebas de estrés están diseñadas para enfrentar al programa a condiciones anormales. La prueba realiza peticiones a un sistema, de manera que demande recursos en cantidad, frecuencia o volúmenes extremos.

Las pruebas fueron realizadas utilizando la herramienta Apache Jmeter. El sistema fue instalado en un entorno de pruebas con las siguientes características:

- Microprocesador Core 2 Duo.
- Memoria RAM de 4 GB.
- Disco duro de 250 GB.
- Sistema Operativo Linux Mint 17.2.

La prueba de carga y estrés fue realizada a través de la herramienta Apache Jmeter, la cual brinda los resultados del rendimiento del sistema a partir de los casos de prueba que se ejecutan, los mismos se enfocan en la ejecución de las peticiones *HTTP* a la aplicación. Para ello se realizó una prueba donde para 1500 usuarios conectados concurrentemente en un intervalo de 1 segundo, se obtuvo lo mostrado en la figura que se muestra a continuación.



Etiqueta	# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
Petición HTTP	1500	34657	29994	66150	3042	107833	8,13%	11,5/sec	6423,4
Total	1500	34657	29994	66150	3042	107833	8,13%	11,5/sec	6423,4

Ilustración 21: Prueba de Carga y Estrés. Reporte resumen

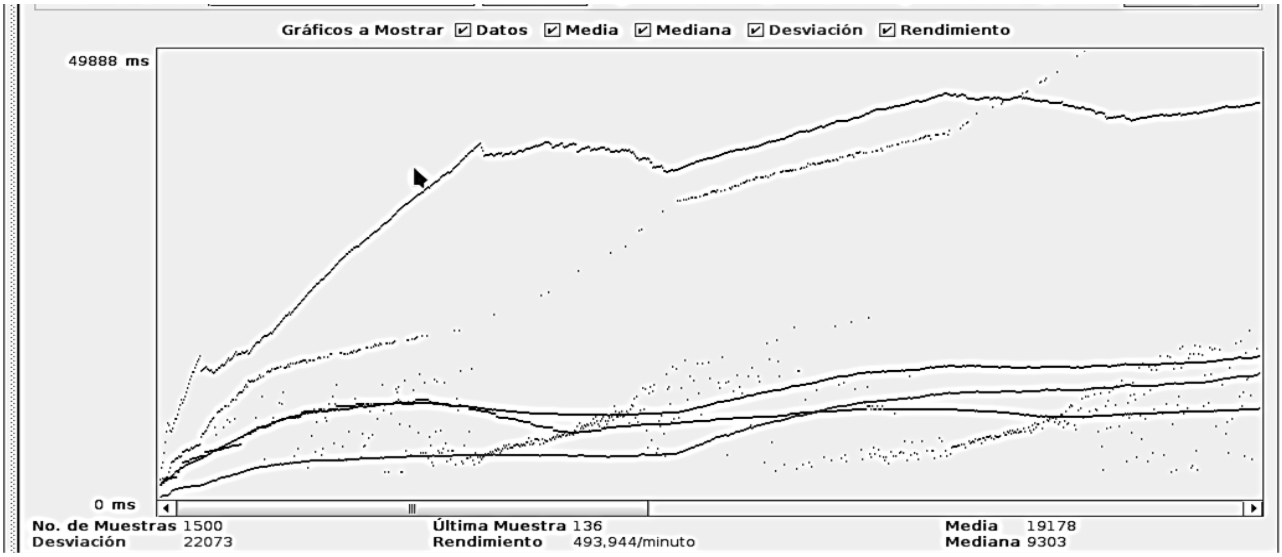


Ilustración 22: Prueba de Carga y Estrés. Gráfico Resumen

Luego de aplicada la prueba se obtuvo un tiempo mínimo de 3 segundos al cargar la página y 107 segundos como tiempo máximo, para un porciento de error de 8,13 %, donde el promedio de peticiones atendidas fue de 11,5 peticiones por segundo. El sistema es capaz de procesar 280 documentos en un tiempo de 2 minutos. Esto evidencia que el sistema puede procesar la carga esperada, cumpliéndose de este modo el requisito no funcional.

3.4.4. Pruebas de Seguridad

La seguridad informática comprende la puesta en práctica de un conjunto de medidas preventivas y reactivas en los sistemas informáticos y tecnológicos, que posibilitan la protección de la información, persiguiendo como objetivo principal la integridad, confidencialidad y disponibilidad de la misma (INTECO-CERT, 2014). El objetivo de esta prueba es evaluar el funcionamiento de los controles de seguridad del sistema para asegurar la integridad y confidencialidad de los datos. El objetivo principal es probar las vulnerabilidades del sistema frente a accesos o manipulaciones no autorizadas.

Para la ejecución de esta prueba se utilizó la herramienta Acunetix la cual se encarga de señalar las vulnerabilidades del subsistema de búsqueda de noticias del buscador Orión. Ejecutada la primera iteración, la herramienta utilizada arrojó los siguientes resultados:

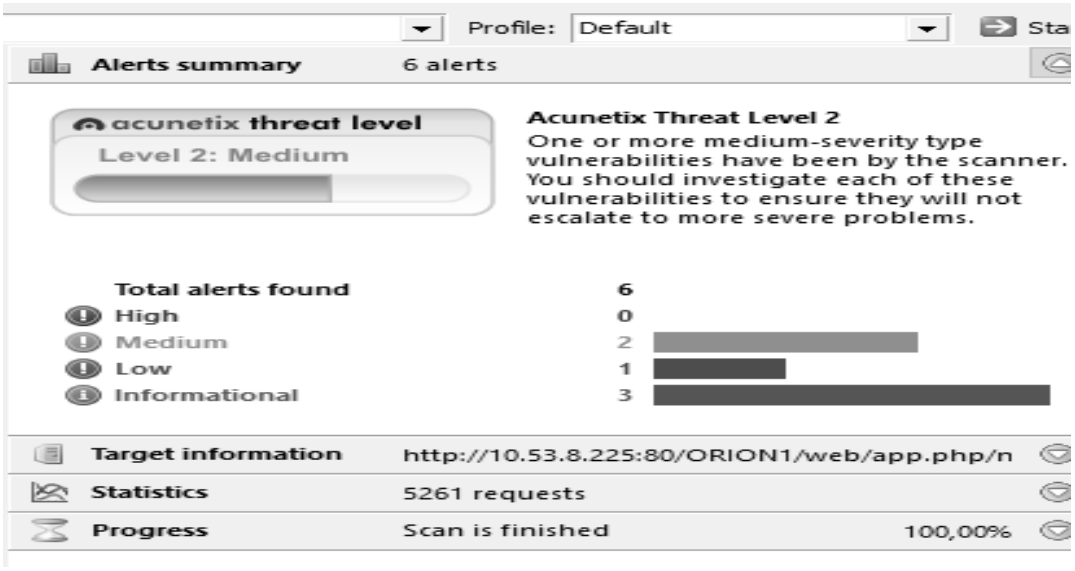


Ilustración 23: Prueba de seguridad (Iteración 1)

Las alertas se clasificaron en:

- Formulario HTML sin protección CSRF.

También conocido como un ataque o sección de manejo de un solo clic y abreviado como CSRF o XSRF, es un tipo de *exploit*[21] malicioso de un sitio web mediante el cual los comandos no autorizados se transmiten de un usuario que confía el sitio web. Acunetix WVS encontró un formulario HTML sin aparente protección CSRF implementado.

- La página de error del servidor muestra demasiada información.

Mediante la solicitud de una página que no existe, se devuelve una página de error. Esta página de error contiene el número de versión del servidor web y una lista de los módulos habilitados en este servidor. Esta información puede ser utilizada para llevar a cabo nuevos ataques.

- Posibles archivos confidenciales.

Acunetix WVS encontró un directorio sensible. Este directorio no está directamente vinculado desde el registro del sitio web, puede ser un recurso sensible, como los directorios de copia de seguridad de base de datos, páginas de administración, directorios temporales, entre otros; cada uno de estos directorios puede ayudar a un atacante para aprender más acerca de su destino.

[21] Secuencia de comandos y/o acciones, utilizada con el fin de aprovechar una vulnerabilidad de seguridad

Luego de realizada una primera iteración y encontrados los errores antes descritos se llevó a cabo medidas para la solución de los mismos. Luego fue ejecutada una segunda iteración dando como resultados cero no conformidades.

3.5. Conclusiones del capítulo

En el presente capítulo se trataron aspectos importantes sobre el desarrollo del subsistema, algunos de estos aspectos fueron el diagrama de componentes y las pruebas de *software*.

- ✓ El diagrama de componentes y sus descripciones permitieron tener una mejor idea de la estructura y el funcionamiento de estos componentes en el sistema.
- ✓ Las pruebas al sistema contribuyeron a la detección de una serie de no conformidades, las cuales fueron solucionadas.
- ✓ Se hizo uso de los estándares de codificación fomentando las buenas prácticas en el desarrollo del subsistema.

Conclusiones generales

Una vez terminada la presente investigación donde los objetivos específicos propuestos fueron cumplidos, y como resultado posibilitó la implementación del Subsistema de búsqueda de noticias del buscador Orión, se desglosan las siguientes conclusiones:

1. La investigación permitió analizar las tendencias actuales de los buscadores y rastreadores, identificando posibles servicios y luego adaptarlos al producto desarrollado.
2. El levantamiento de requisitos permitió identificar las funcionalidades y servicios implementados en el Subsistema de búsqueda de noticias del buscador Orión.
3. La elaboración de los artefactos propuestos por la metodología de desarrollo permitió una mayor comprensión del sistema a desarrollar, así como la identificación de los procesos y características del mismo.
4. Las pruebas de software realizadas, permitieron detectar no conformidades, las cuales fueron resueltas, garantizando la calidad del producto al comprobar su correcto funcionamiento.

Recomendaciones

A partir de las experiencias obtenidas en el desarrollo del trabajo de diploma y con el fin de adquirir un aprovechamiento óptimo del resultado alcanzado se recomienda:

- Realizar un mejoramiento de la interfaz teniendo en cuenta los estándares actuales.
- Realizar un *plugin* que recupere noticias vía html el cual necesitará uso de inteligencia artificial para la identificación de las categorías y discriminación de los artículos para conocer si son noticias o no.

Referencias

Arias Molina, Angela. 2011. Rincondelvago. [En línea] 2011. http://html.rincondelvago.com/internet-como-un-nuevo-medio-de-comunicacion.html.

Belén Fuentes, Melisa. 2013. 3 de julio de 2013, suite101.

Bonilla, GRAL. JUAN CRISÓSTOMO. 2014. BUSCADORES DE INTERNET. [En línea] 2014. [Citado el: 27 de Octubre de 2015.] https://docs.google.com/document/d/1w75wqjdGmRmLmic9_nFC3pm1yUogCgTEFv26VcQaEKE/edit?pli=1.

Creative Commons. 2014. CCM. *CCM.* [En línea] 2014. http://es.ccm.net/contents/304-lenguajes-de-programacion.

Crovi Druetta, Dra. Delia María. 2006. ¿ES INTERNET UN MEDIO DE COMUNICACIÓN? [En línea] 10 de Junio de 2006. [Citado el: 23 de Octubre de 2015.] http://www.revista.unam.mx/vol.7/num6/art46/jun_art46.pdf.

Cuadrado, Enrrique Mauricio Durán. 2012. ESTUDIO SOBRE EL USO DE LA TÉCNICA PERIODÍSTICA EN EL GÉNERO DE LA NOTICIA. [En línea] Febrero de 2012. dspace.ups.edu.ec/bitstream/123456789/3529/1/QT03154.pdf.

definicion.de. 2008. definicion.de. [En línea] 2008. http://definicion.de/buscador-2/.

Domínguez Goya, Emelia . 2012. Medios de comunicación masiva. [En línea] 2012. http://www.aliat.org.mx/BibliotecasDigitales/comunicacion/Medios_de_comunicacion_masiva.pdf.

Ecured. **2013.** 2013.

Gil, Manuel Torres. 2016. Universidad de Almeria. *Universidad de Almeria.* [En línea] 5 de Abril de 2016. http://indalog.ual.es/mtorres/LP/index.php?opcion=inicio.

Junta de Andalucía. Convenio de codificación específico para Drupal. 2013. Marco de desarrollo de la Junta de Andalucia. *Marco de desarrollo de la Junta de Andalucia.* [En línea] 4 de Abril de 2013. http://www.juntadeandalucia.es/servicios/madeja/contenido/libro-pautas/8.

Los buscadores más populares de Internet. **Elizalde, Erika.** about.

Marín Ochoa, Beatriz Elena. 2009. La infografía digital, una nueva forma de comunicación. [En línea] 15 de noviembre de 2009. http://www.tdx.cat/bitstream/handle/10803/48653/bemo1de1.pdf?sequence=1.

Martínez Castro, María Estela , Cárdenas Olivares, Gabino y Banderas Martínez, Cuauhtémoc . 2013. EL LENGUAJE LITERARIO EN EL PERIODISMO: TRAZOS Y RETAZOS. [En línea] 2013. http://congreso.pucp.edu.pe/alaic2014/wp-content/uploads/2013/10/vGT16-Mart%C3%ADnez-C%C3%A1rdenas-Banderas.pdf.

Martínez Méndez, Francisco Javier. 2004. *RECUPERACIÓN DE INFORMACIÓN: MODELOS, SISTEMAS Y EVALUACIÓN.* 2004.

Rodríguez, Alfonso Pérez. 2014. *Estructuración y Especificación de Casos de Uso.* 2014.

RUIZ REY, FRANCISCO JOSÉ. WEB 2.0. UN NUEVO ENTORNO DE APRENDIZAJE EN LA RED. [En línea] http://dim.pangea.org/revistaDIM13/Articulos/pacoruiz.pdf.

Sistemas de acceso a la información de prensa digital: tipología y evolución. **Guallar, Javier , Abadal, Ernest y Codina, Lluís . 2013.** 61, 12 de abril de 2013, Vol. 27.

Sun Microsystems Inc. 1999. *Convenciones de código para el lenguaje de programación Java.* s.l. : Sun Microsystems Inc., 1999.

Travieso Aguiar, Ing. Mayelín. 2008. Las publicaciones electrónicas: una revolución en el siglo XXI. [En línea] 5 de mayo de 2008. http://www.bvs.sld.cu/revistas/aci/vol11_2_03/aci010203.htm.

turismotecnologico. [En línea] http://turismotecnologico.bligoo.com.mx/media/users/35/1750075/files/661142/Buscadores.pdf.

Una breve historia de Internet. **Otros, Leiner Barry M. 2010.** 2010.

CABRERA GONZÁLEZ, L. y POMPA TORRES, E. R. *Extensión de Visual Paradigm for UML para el Desarrollo Dirigido por Modelos de aplicaciones de gestión de información.* 2012. 80 págs. *h*http://repositorio_ institucional.uci.cu//jspui/handle/ident/TD_05815_12*i*

[BAE, 1992] Baeza-Yates, R. and Frakes, W.B. Information retrieval : data structures & algorithms Englewood Cliffs, New Yersey : Prentice Hall, 1992 504 p. ISBN 0-13-463837-9

[BAE, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York :ACM Press ; Harlow [etc.] : Addison Wesley, 1999 XX,513 p. ISBN 0-201-39829-X

Baeza- Yates, Ricardo. 2008. "Cómo funciona la Web," August 8.

"Banrepcultural." 2015. http://www.banrepcultural.org/blaavirtual/ayudadetareas/comunicacion/los_medios_de_comunicacion#Medios_digitales.

Bello, István Ojeda. 2012. "¿Qué Es Noticia? - Monografias.com." http://www.monografias.com/trabajos35/la-noticia/la-noticia.shtml.

Buscadores Web. 2015. "Ranking Buscadores Internet." *Buscadores Web*. http://buscadores-web.com/ranking-buscadores-internet/.

Cabrera, Humberto Cardoso, and Yaudel Castillo Pérez. 2013. "Bing - EcuRed." http://www.ecured.cu/Bing.

Cabrera Gerra, Ivis, Roexcy Vega Prieto, and Guillermo Báez Ramos. 2009. "Propuesta de Un Modelo Base Para Un Sistema de Busqueda de Videos Digitales a Traves de Metadatos." http://repositorio_institucional.uci.cu//jspui/handle/ident/TD_2756_09.

Cabrera González, Lianet, and Enrique Roberto Pompa Torres. 2012. "Extensión de Visual Paradigm for UML Para El Desarrollo Dirigido Por Modelos de Aplicaciones de Gestión de Información." http://repositorio_institucional.uci.cu//jspui/handle/ident/TD_05815_12.

Castro, María Estela Martínez, Gabino Cárdenas Olivares, and Cuauhtémoc Banderas Martínez. 2014. "EL LENGUAJE LITERARIO EN EL PERIODISMO: TRAZOS Y RETAZOS." http://congreso.pucp.edu.pe/alaic2014/wp-content/uploads/2013/10/vGT16-Mart%C3%ADnez-C%C3%A1rdenas-Banderas.pdf.

"Conozca Más Sobre La Tecnología Java." 2016. Accessed May 26. https://www.java.com/es/about/.

Craig, larman. 2003. *UML Y Patrones*. 2da Edicion. http://sunshine.prod.uci.cu/book/5183103e0571741ed7000060/.

Croft, W. Bruce. 2005. *: Ebook Charting a New Course Natural Language Processing and Information Retrieval*. http://sunshine.prod.uci.cu/gridfs/sunshine/books/ebook_-_Charting_a_New_Course_-_Natural_Language_Processing_and_Information_Retrieval.pdf.

CUBA. 2015. "Manual de Usuario." http://www.redcuba.cu/ayuda/ayuda.pdf.

"Diagramas de Clases." 09:03:08 UTC. http://es.slideshare.net/jpbthames/diagramas-de-clases.

Dolores. 2011. "Historia de Los Buscadores | Historia de La Informática." http://histinf.blogs.upv.es/2011/01/11/1808/.

"Elastic." 2016. Accessed February 9. https://www.elastic.co/guide/en/elasticsearch/guide/current/intro.html.

elastic.co. 2016. "Getting Started." Docs/Elasticsearch/Definitive Guide/2.x. Accessed April 21. https://www.elastic.co/guide/en/elasticsearch/guide/current/getting-started.html.

Exite. 2015. "Excite España." http://info.excite.es/.

Fernández, Nadala. 2008. "Historia de Internet." http://www.fib.upc.edu/retro-informatica/historia/internet.html.

"Heritrix - Home Page." 2016. Accessed February 9. http://crawler.archive.org/index.html.

Lucene.apache.org. 2015. "Apache Solr -." http://lucene.apache.org/solr/index.html.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2018. "An Introduction to Information Retrieval," July 12. http://sunshine.prod.uci.cu/gridfs/sunshine/books/Information_Retrieval.pdf.

Mark Otto, and Jacob Thornton. 2016. "Bootstrap 3, El Manual Oficial." Accessed May 26. https://librosweb.es/libro/bootstrap_3/.

McDonald, Sharon. 2008. *Advances in Information Retrieval*. http://sunshine.prod.uci.cu/gridfs/sunshine/books/ebook_-_Advances_in_Information_Retrieval.pdf.
MDN. 2016. “HTML.” *Mozilla Developer Network*. Accessed May 26. https://developer.mozilla.org/es/docs/Web/HTML.
Medina García, Anay, Meylin Martínez Chong, and Yusniel Hidalgo Delgado. 2011. “Propuesta Arquitectónica de Un Sistema de Recuperación de Información Geográfica Para El Motor de Búsqueda Orión.” http://repositorio_institucional.uci.cu//jspui/handle/ident/TD_04252_11.
“Medios de comunicación masiva.” 2015. Accessed November 23. http://www.portaleducativo.net/septimo-basico/317/Medios-de-comunicacion-masiva.
microsoft. 2016. “Información General Sobre ASP.NET.” Accessed May 21. https://msdn.microsoft.com/es-es/library/4w3ex9c2(v=vs.100).aspx.
Monografias.com, silver. 2016. “Un Sitio En Internet. Qué Es La Web? - Monografias.com.” Accessed May 12. http://www.monografias.com/trabajos5/laweb/laweb.shtml#I4.
Nugraha, Arie. 2014. “Indexing Bibliographic Database Content Using MariaDB and Sphinx Search Server.” *The Code4Lib Journal*, no. 25(July). http://journal.code4lib.org/articles/9793.
NutchWiki. 2015. “NutchTutorial - Nutch Wiki.” https://wiki.apache.org/nutch/NutchTutorial.
php.net. 2016. “PHP: ¿Qué Es PHP? - Manual.” Accessed May 21. http://php.net/manual/es/intro-whatis.php.
Pinto Molina, Maria. 2011. “BUSQUEDA Y RECUPERACIÓN DE INFORMACIÓN.” April 13. http://www.mariapinto.es/e-coms/recu_infor.htm#ri2.
Pressman, Roger S. 2005. “Ingeniería del Software.” March 9. http://sunshine.prod.uci.cu/book/4e6a955d05717458640000e8/.
Pupo Gómez, Yordanka, Grabiel Montero Coello, Deivis Ricardo Álvarez Mendoza, and Aliennis Mercedes González Hurtado. 2014. “HERRAMIENTA INFORMÁTICA PARA LA ELABORACIÓN DE MAPAS CONCEPTUALES COLABORATIVOS BASADOS EN LA WEB 2.0,” October. http://repositorio_institucional.uci.cu//jspui/handle/ident/8135.
python.org. 2016. “Welcome to Python.org.” *Python.org*. https://www.python.org/.
Quintana, Eduardo. 2016. “IDC: Habrá 3.200 Millones de Usuarios de Internet En 2016 » MCPRO.” *MuyComputerPRO*. January 1. http://www.muycomputerpro.com/2016/01/01/idc-habra-3-200-millones-de-usuarios-de-internet-en-2016.
Ramos, Elizabeth Cejudo. 2013. “Mujer, periodismo y opinión pública en Sonora: el caso de los periódicos El Pueblo y El Tiempo de Hermosillo.” http://www.scielo.org.mx/pdf/regsoc/v26n61/v26n61a13.pdf.
“Revistero Virtual - Ayuda - ¿Qué Es Una Publicación Digital?” 2016. Accessed April 28. https://revisterovirtual.com/help_sub_topics/39/ayuda/%BFQu%E9+es+una+publicaci%F3n+digital%3F.html.
Rey, Fransisco José Ruiz. 2013. “WEB 2.0. UN NUEVO ENTORNO DE APRENDIZAJE EN LA RED.”
Rovira, Cristòfol, Jofre Capdevila, and Mari Carmen Marcos. 2014. “La Importancia de Las Fuentes En La Selección de Artículos de Prensa En Línea: Un Estudio de Google Noticias Mediante Seguimiento Ocular (Eye-Tracking).” *Investigación Bibliotecológica* 28 (63): 15–28.
Rueda, Eyeris Rodriguez, and Yusniel Hidalgo Delgado. 2012. “LOS SPIDERS Y SU FUNCIÓN EN LOS MOTORES DE BÚSQUEDA,” March 12. http://repositorio_institucional.uci.cu/jspui/handle/ident/4130.
SearchEngineLand. 2015a. “Search Engine Land’s Guide To Yahoo.” *Search Engine Land*. http://searchengineland.com/guide/yahoo.
———. 2015b. “The B2B Marketer’s Guide to Baidu SEO.” *Search Engine Land*. http://searchengineland.com/the-b2b-marketers-guide-to-baidu-seo-180658.
Senn, James A. 1996. *Análisis Y Diseño de Sistemas*. Mc Graw Hill, México. http://wwwisis.ufg.edu.sv/wwwisis/documentos/TE/025.04-F634d/025.04-F634d-BG.pdf.
SlideShare. 03:04:28 UTC. “Metodologias de Desarrollo.” http://es.slideshare.net/MeneRomero/metodologias-de-desarrollo.

———. 2015. “Motores de Búsqueda AOL.” http://es.slideshare.net/linamarcela0607/motores-de-bsqueda-aol-44880864.
Solana, Luis. 1988. “Medios de Comunicación Social Y Telecomunicaciones.” *Cuenta Y Razón*, no. 34: 17–20.
SOMMERVILLE. 2005. *Ingeniería Del Software*. 7ma ed.
sunshine. n.d. *Secretos de los Buscadores*. http://sunshine.prod.uci.cu/gridfs/sunshine/books/Secretos_de_los_Buscadores.pdf.
Torres Pombert, Lic. Ania. 2003. “El Uso de Los Buscadores En Internet.” http://bvs.sld.cu/revistas/aci/vol11_3_03/aci04303.htm.
TwigHomepage. 2016. “Homepage - Twig - The Flexible, Fast, and Secure PHP Template Engine.” Accessed May 26. http://twig.sensiolabs.org/.
UPREDES. 2015. “LUPA - Acerca de.” Accessed November 27. http://lupa.upr.edu.cu/acerca-de.html.
“Visual Paradigm Essential.” 2016. *Udemy*. Accessed February 9. https://www.udemy.com/visual-paradigm-essential/.
“WIRE (Web Information Retrieval Environment)::Center for Web Research.” 2015. Accessed November 12. http://www.cwr.cl/projects/WIRE/.
“YAML Ain’t Markup Language.” 2016. Accessed May 26. http://www.yaml.org/about.html.