Universidad de las Ciencias Informáticas

FACULTAD 6

Título: Sistema recuperador de noticias digitales.

Trabajo de Diploma para optar por el título de

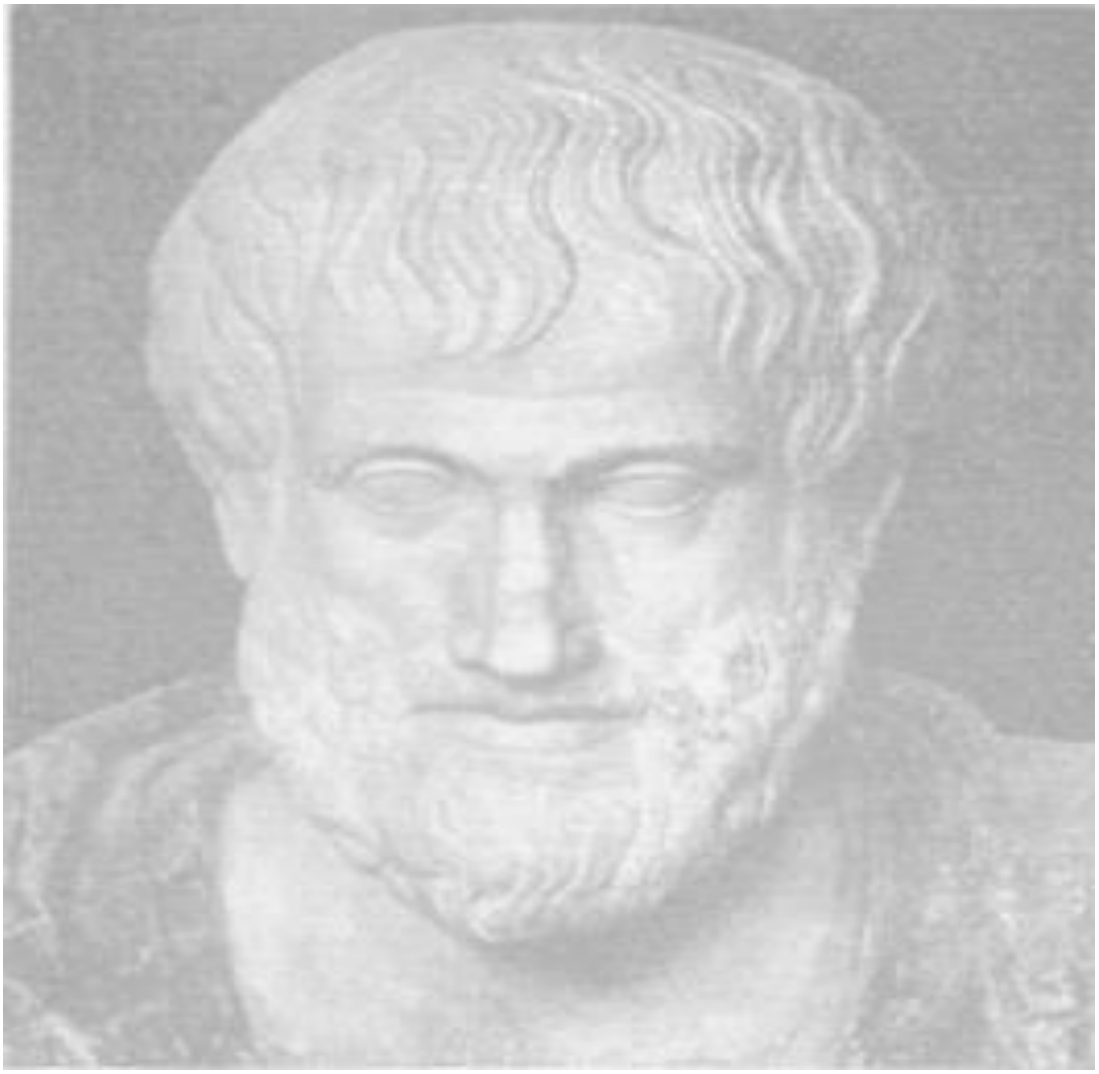
Ingeniero en Ciencias Informáticas

Autor: Marcial Hinojosa Esteva.

Tutor: Ing. Carlos de Jesús Andrés González.

"13 de junio de 2013"

Frase



La inteligencia consiste no sólo en el conocimiento, sino también en la destreza de aplicar los conocimientos en la práctica.

Aristóteles

Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

<nombre autor>

_____

Firma del Autor

<nombre tutor>

_____

Firma del Tutor

Datos de contacto

Generales del diplomante:

Nombre y apellidos: Marcial Hinojosa Esteva.

Sexo: Masculino.

Grupo: 6405.

Correo: mhinojosa@estudiantes.uci.cu.

Generales del tutor:

Tutor: Ing. Carlos de Jesús Andrés González.

Categoría docente: Especialista General.

Centro de trabajo: Universidad de las Ciencias Informáticas.

Título de la especialidad de graduado: Ingeniero en ciencias Informáticas.

Año de graduación: 2010.

Institución en la que se graduó: Universidad de las Ciencias Informáticas.

Correo electrónico: candres@uci.cu.

Teléfono particular: 837 3093.

Dedicatoria

A mi esposa Lisandra, por estar conmigo siempre en las buenas en las malas, por darme cariño, apoyo y confianza en cada segundo, cada instante que viví en esta escuela, y, sobre todo, por darme lo más bello que alguien en el mundo puede tener, una bebé.

A mi hija Vanessa que, aunque no ha nacido todavía, la amo desde el primer instante que se formó en la pancita de su mami. Es lo más grande que me ha pasado en mi vida y quiero que se siente orgullosa de mí.

A mis padres Marcial y Carolina por apoyarme siempre y guiarme por el sendero correcto de una forma u otra. Ustedes han hecho realidad este sueño.

A mi abuelita querida Pura la cual quiero mucho y es la persona que siempre me complace en todo sin importar que sea, la persona que siempre me dio ánimo para llegar hasta donde estoy hoy.

A mi abuelo Rene que lo quiero muchísimo y siempre estuvo ahí para mí cuando lo necesite y el cual considero como un padre más.

A mi hermano Andy que aunque muy joven, brindó su apoyo también para que este día hoy llegara. Para que le sirva de ejemplo como hermano mayor de que no hay metas imposibles en la vida.

A todos los demás familiares que de una manera u otra mostraron su preocupación por mi carrera e influyeron en el transcurso de la misma.

A todos mis amigos y amigas que considero como hermanos.

Agradecimientos

A Dios primeramente por haberme dado la familia que tengo hoy en dia y por haberme concedido el más bello deseo que un padre pueda pedir.

A mi madre Carolina querida que la que me trajo al mundo y creo en mí ese espíritu de sacrificio y lucha por lograr lo que quiero. Te quiero mucho mami.

A mi padre Marcial, que aunque no vive conmigo, su apoyo y cariño nunca falto y siempre sembró el estado de ánimo en mí.

A mi madrastra Gladys que, aunque no es mi madre, se comportó como una y me dio confianza y ánimo para llegar hoy a donde estoy.

A mis abuelitos queridos que lo son todo para mí, que siempre me ayudaron en todo, me lo dieron todo y se sacrificaron tanto por mí, los adoro con el alma.

A mi esposa Lisandra, mi más cercana compañera y mi más grande amor. Y a su familia de por darme su apoyo incondicional siempre y tratarme como un hijo y un integrante más de la familia.

A mío tío Alfredo y su familia por estar siempre presente y disponibles para brindarme su apoyo.

A mi tutor Carlos, por siempre estar ahí encima de mí, preocupado por el desarrollo de la tesis, siendo meticuloso por cada detalle para que todo fluyera con éxito.

A mis amigas Liz y Yamila por estar conmigo ahí 5 años en los momentos malos y buenos y por brindarme su amistad. A ustedes, gracias.

A mis amigos del apartamento que hoy están, Eliober, Edgar, Falcón, Alejandro, Luis Manuel, y los que no también, gracias por su amistad y su compañía todo este tiempo. Al igual que las muchachitas de mi aula que son divinas, a todas ellas, gracias por dejarme ser su amigo.

A todos que aquellos que alguna vez se preocuparon por mí y me preguntaron por la tesis. A todos, gracias.

Resumen

Con el desarrollo de las tecnologías de la información y las comunicaciones (TIC) y el surgimiento de Internet, la recuperación de noticias se ha convertido en la principal tarea para las agencias que guían su trabajo hacia el proceso de divulgación de la información. Las entidades o instituciones que se dedican a este tipo de funcionalidades requieren de mecanismos de búsqueda que posibiliten la recuperación precisa de la noticia y todos sus componentes asociados.

En Cuba existe la Agencia de Información Nacional (AIN) entidad que se basa en la combinación de tecnologías para dar como fruto la divulgación de informaciones a través de diversos canales. En ella se utilizan una herramienta de vital importancia denominada winKertin (Yunior, 2008), utilizada por el redactor a la hora de la recuperación de la información, sin embargo la misma no permite la obtención completa de la noticia, excluyendo algunos de los elementos que la componen.

El presente trabajo constituye la realización de un sistema de recuperación y extracción completa de las noticias publicadas en Internet, permitiendo su posterior almacenamiento en un sistema gestor de base de datos, dando cumplimiento a los requisitos funcionales de la aplicación y garantizando un despliegue fácil de la aplicación. En su elaboración se han utilizado tecnologías completamente libres que permiten dar respuestas a las exigencias del país y constituye una alternativa factible y aplicable en cualquier entidad que necesite recuperar noticias digitales.

Palabras Clave: AIN, búsqueda, extracción, noticias, recuperación, winKertin.

Tabla de Contenidos

Introducción

La razón de ser de las agencias de noticias hay que buscarla en su origen, ya que tienen como objetivo solventar una necesidad. Por lo tanto, ya desde el primer momento, las agencias de noticias parten como un servicio a los medios de comunicación, con un papel de intermediario entre los acontecimientos y la empresa informativa que, finalmente, hará efectiva la difusión de las noticias. De acuerdo con el diccionario general de periodismo de José Martínez de Sousa, la agencia es "una empresa que presta determinados servicios". Por lo tanto, una agencia de noticias es un sistema de recolección de noticias que distribuye regularmente sus servicios informativos entre diversos medios de comunicación suscritos a los mismos.

Hoy en día a nivel mundial existen diversas agencias de noticias que se caracterizan por mantener una extensa red de corresponsales en todo el mundo y por la posibilidad de difusión de sus servicios, que cubren la mayor parte del planeta, entre ellas se destacan Associated Press (AP), Agence France-Presse (AFP), Reuters (REU), Deutsche Presse-Agentur (DPA), entre otras. En Cuba existe la Agencia Cubana de Noticias (ACN), división de la Agencia de Información Nacional (AIN), la cual tiene como objetivo principal difundir el acontecer noticioso nacional como internacional, apoyándose en diferentes fuentes informativas como pueden ser periódicos nacionales, agencias de prensas y sitios Web. La ACN posee varios medios para la difusión de los contenidos informativos, entre ellos una radio, un canal de televisión y un sitio web, además cuenta con un banco de fotos y un cast[1] de noticia. Los procesos de recuperación de las noticias y su consecuente divulgación a través de los diversos medios existentes en la AIN, cuentan con dos vías para dar cumplimiento a dicha función. Una primera vía y la principal, está asociada a la utilización de una herramienta de recuperación la cual se encarga de obtener la información noticiosa en formato texto. Como segunda vía se emplea el uso de los motores de búsquedas para la recuperación de noticias que solo presentan texto así como la obtención de audio, video o imagen. Ambos mecanismos son independientes por lo que el proceso de redacción de la noticia en la entidad no se realiza con la mayor eficiencia, influyendo esto en el desarrollo ágil y eficiente en la gestión de los procesos de la información y en la necesidad de la misma para el usuario.

[1] Reparto, conjunto.

Teniendo en cuenta los objetivos de la entidad y la necesidad de mantener informado al usuario, se ha determinado el siguiente **problema a resolver**, ¿cómo lograr la completa obtención de la noticia y sus elementos asociados, para su posterior procesamiento y difusión por los medios de la AIN? Como **objeto de estudio** para la realización del presente trabajo se tomaron en cuenta los procesos de recuperación, extracción y almacenamiento de las noticias publicadas en Internet. Su **campo de acción** está centrado en la creación de una aplicación que permita la obtención automática de la noticia y sus elementos asociados para la AIN. Se planteó como **objetivo general** del trabajo**:** desarrollar una aplicación que permita automatizar la recuperación, extracción y almacenamiento de las noticias y sus componentes asociados para la AIN. Para dar cumplimiento al objetivo general se trazaron los siguientes **objetivos específicos:**

1. Definir la tecnología a emplear para la recuperación de noticias digitales.
2. Definir el procedimiento para la extracción de noticias digitales.
3. Definir una estructura para almacenar la noticia y sus elementos asociados.
4. Implementar el sistema recuperador de noticias digitales.
5. Realizar pruebas al sistema recuperador de noticias digitales.

Para guiar la investigación se planteó la siguiente **idea a defender**: Si se implementa un sistema de recuperación, extracción y almacenamiento de las noticias y sus elementos asociados para la AIN, entonces se garantizará los procesos de gestión de la noticia llevado a cabo por los editores de la agencia.

Para dar cumplimiento al objetivo general se definieron las siguientes **tareas de la investigación:**

1. Caracterización de los formatos de redifusión web RSS y Atom.
2. Caracterización de buscadores y metabuscadores que permitan la interacción mediante servicios.
3. Identificación de la(s) tecnología(s) a emplear para la recuperación de noticias.
4. Caracterización de los algoritmos que permitan la extracción del contenido de una noticia a partir de un archivo HTML.
5. Evaluación de los algoritmos identificados anteriormente y proponer la utilización de uno de ellos o uno de propia elaboración.
6. Caracterización de la estructura de las noticias y los elementos que la componen.
7. Definición de una estructura de almacenamiento para guardar las noticias recuperadas de Internet.

8. Implementación de una aplicación que permita la recuperación, extracción y almacenamiento de la noticia y sus contenidos asociados.
9. Realización de pruebas del funcionamiento a la aplicación desarrollada.

Se utilizaron diferentes métodos que permitieron obtener información valiosa para el desarrollo de la investigación:

Métodos teóricos.

Analítico – Sintético: Aportó la posibilidad de buscar la esencia de los fenómenos, los rasgos que caracterizan y distinguen los procesos relacionados con el análisis de la redacción de noticias de la Agencia de la Información Nacional. Dentro de la investigación ha permitido la extracción de los elementos más importantes que se relacionan con la entidad.

Inductivo – Deductivo: Este método posibilita el desarrollo de un conocimiento más generalizado del análisis de los procesos de la redacción de noticias de la Agencia de la Información Nacional, partiendo de aspectos específicos documentados en investigaciones del tema.

El trabajo de diploma está estructurado de la siguiente forma:

Capítulo 1. Fundamentación teórica, durante este capítulo se enuncian los principales conceptos teóricos asociados a la investigación, los cuales posibilitan un mejor entendimiento de la situación problemática planteada y el marco del problema en sentido general.

Capítulo 2. Análisis y presentación de la solución propuesta, mediante este capítulo se realiza el modelo de dominio concerniente al problema planteado, así como un levantamiento de los requisitos de software. Se describe la propuesta de solución a través de los actores, casos de uso y la descripción de cada uno de los caso de uso.

Capítulo 3. Construcción y evaluación de la solución, en este capítulo se describen los diagramas de clases y principios del diseño, la arquitectura del sistema, los elementos sobre la base de datos a utilizar, así como el diagrama de despliegue y el modelo de implementación.

Capítulo 4. Validación de la Solución Propuesta. En este capítulo se realizan las pruebas de caja negra a la solución creada en el capítulo anterior.

Capítulo 1: Fundamentación Teórica.

1.1 Introducción.

En este capítulo se puntualizan algunos de los conceptos fundamentales que permitirán un mejor entendimiento del dominio del problema. De igual modo se realiza un análisis de la situación problemática para definir el sistema recuperador de noticias digitales para la AIN.

1.2 Situación Problemática.

El grupo de periodistas de la AIN, encargados de la obtención de noticias, utilizan diversas vías de comunicación para la recolección de las noticias externas, es decir las noticias elaboradas por otras fuentes de información no pertenecientes a la agencia, para esto se emplea fundamentalmente la utilización de una herramienta especializada en la recepción y transmisión integrada de noticias denominada winKertin, que brinda únicamente la noticia en formato texto, otras de las vías está vinculada con los motores de búsqueda de Internet, entre los que cabe mencionar Yahoo, Altavista, Bing, y principalmente, Google, los cuales facilitan la obtención de un listado de páginas web que contienen información sobre un tema de interés.

Una vez que el winKertin ha cumplido su objetivo, los redactores de noticias deben gestionar las mismas para utilizarlas en los procesos de difusión que se llevan a cabo en la AIN a través de los diferentes medios existentes en la entidad, fundamentalmente el canal de televisión “Señal ACN”, y la emisora de radio “Radio ACN”. Para el caso de la televisión se cuenta, entre otros elementos, con un video para divulgar la noticia, a diferencia de la radio, en la cual solo se necesita mayormente, un resumen de la información obtenida. Por otro lado, las formas de representación que poseen las noticias obtenidas como resultado del funcionamiento de la herramienta de recuperación de información empleada, en este caso el propio winKertin, no satisfacen las necesidades de los lectores a la hora de interactuar con el contenido digital, pues como se mencionó anteriormente las noticias recuperadas solo poseen texto. Es por ello que, los editores de dichas noticias se ven obligados entonces a consultar manualmente los motores de búsqueda que permitan descargar imágenes, o acceder al banco de fotos de la agencia o algún dispositivo físico conectado a la estación de trabajo para posteriormente acceder a las imágenes o videos asociados de acuerdo a la información obtenida en formato texto, para entonces confeccionar la noticia completa.

Entre las características deseadas para el producto, está la obtención de funcionalidades genéricas fácilmente escalables, que no dependan de un entorno dado y no atadas a un diseño gráfico específico; así como la necesidad de que el mismo sistema permita la obtención de la noticia completa, siendo

esto un factor determinante para el redactor encargado de la elaboración de la noticia a la hora del cumplimiento de su función y respondiendo así a las necesidades del cliente.

1.3 Conceptos asociados al dominio del problema.

Los conceptos asociados al problema que se explican a continuación, hacen de este epígrafe, una parte fundamental para el desarrollo del trabajo científico; la información adquirida en la investigación servirán de base para los resultados esperados.

1.3.1 Noticia.

El género informativo tiene como objetivo informar y captar al lector por el camino de la comunicación de las noticias y de la hábil exposición de las ideas, lo que se consigue por medio de diferentes caminos que, a su vez, dan origen a diferentes modalidades del estilo periodístico. Los géneros puros pertenecientes al género informativo son la noticia y la crónica. Dentro de los géneros informativos la noticia ocupa un lugar destacado. (Shabb, 2004).

"La noticia es un hecho verdadero, inédito o actual, de interés general, que se comunica a un público que pueda considerarse masivo, una vez que ha sido recogido, interpretado y valorado por los sujetos promotores que controlan el medio utilizado para la difusión". (Martinez, 1962).

La noticia consiste en la escueta enumeración de los datos de un acontecer a producirse y producido. También se puede considerar a la noticia como el género periodístico por excelencia que da cuenta, de un modo personal, pero completo, de un hecho actual o actualizado, digno de ser conocido y divulgado, y de innegable repercusión humana.

La noticia es el relato objetivo de un suceso cuyo conocimiento importa hacer público oportunamente. (Martinez, 1962). Para que su contenido sea completo y efectivo, debe responder las siguientes preguntas (las 6 W):

1. Quién: El/la protagonista de la noticia.
2. Qué: El suceso.
3. Cuándo: El tiempo.
4. Dónde: El lugar del hecho.
5. Por qué: Las causas.
6. Para qué: Los objetivos.

1.3.2.1 Estructura de la noticia.

La estructura de las noticias es un factor fundamental en el entendimiento del texto por parte de los lectores y del análisis de hechos noticiosos. Cuando se escriba una noticia se comienza siempre por lo más importante. Los datos se van distribuyendo a lo largo de la noticia por el grado de interés que tengan. Este esquema se conoce en la profesión como la estructura de la pirámide invertida y pretende cumplir dos objetivos: el primero y más importante es que de esta forma el lector puede informarse de lo más importante de la noticia con rapidez, si por cualquier motivo interrumpe la lectura en el cuarto o quinto párrafo se habrá enterado de los aspectos más importantes referidos a ese acontecimiento. Si prosigue su lectura, podrá completar su información enterándose de más matices y profundizando sobre el acontecimiento. (Martinez, 1962).

Figura 1. Estructura de la noticia.

1. El título**:** Tiene la misión de proporcionar lo esencial de la información. Pero tiene también por objetivo suscitar el interés del lector, invitándolo a leer la noticia. De ahí su importancia. (Martinez, 1962). En el sistema se contará con un campo denominado "Titulo" para cada noticia.
2. El subtítulo**:** es una síntesis de lo más importante del texto, por lo que debe ser llamativa. (Martinez, 1962). Este campo se reflejara en el sistema aunque es probable que no en todas las páginas se refleje debido a que no hay una manera común para su identificación.

3. Resumen**:** Es el primer párrafo o unas líneas iniciales en que se resume lo esencial del hecho noticioso. Su redacción responde a las cinco preguntas fundamentales: quién, qué, cuándo, dónde y por qué. Puede faltar alguno de tales elementos o añadirse otros: para qué, cómo. (Martinez, 1962). Se hará presencia de un campo denominado “Resumen” en el sistema ya que, a la hora de la publicación de la noticia, da una síntesis al lector del tema a instruirse.
4. El cuerpo de la noticia: entre varios párrafos, es norma que estos se sucedan siguiendo un orden decreciente, pueden ser más o menos largo y agregar más o menos detalles a lo dicho en la entradilla. Cuando consta de importancia se tiene en cuenta que: si los límites de espacio lo exigen, a la hora de componer la página puede prescindirse de uno o más párrafos empezando por el final, de modo que sólo se supriman los datos menos relevantes. (Martinez, 1962). De la misma forma se reflejara un campo denominado “Cuerpo” el cual tendrá toda la información más importante del contenido de la noticia.

Según (Martinez, 1962). Esta es la estructura que representa el concepto de noticia, sin embargo, dada la era digital que se vive y el desarrollo día a día de la misma así como el desarrollo de las tecnologías que la sustentan y de la tendencia actual, se decidió tener constancia de otros elementos que son importante para el redactor a la hora de realizar el proceso de divulgación de la noticia. Ellos son:

5. Elementos asociados: no es más que aquellas medias con las que cuenta el contenido de la noticia, dígase audio, video o imagen, como se ilustra en la **Figura 1**. Se contará en el sistema con un campo denominado “Elementos asociados” el cual contendrá, de existir en la página web, cada una de estas medias.
6. La Fuente a la que pertenece una noticia: hoy en día se identifica mayormente con la url de la noticia, o sea, el sitio web de donde procede dicha noticia.
7. Fecha de publicación: se refiere al día, mes, año y hora actual en que fue redactada la noticia.
8. Autor: se refiere a la persona que redactó la noticia.

Todos estos son elementos referentes a la estructura de la noticia que se tendrá en cuenta en la presente investigación. En el caso del Autor, conjuntamente con el elemento Subtitulo y con las medias audio y video, mencionadas anteriormente como parte de la estructura de la noticia, su recuperación varía en dependencia de como este estructurado el documento HTML que se esté procesando en ese momento y por consecuente de su aparición o no en el sistema.

1.3.2 Recuperación de la Información.

El tamaño de la web es imposible de medir exactamente y muy difícil de estimar. Sin embargo, se calcula que son decenas de exabytes de información, y crece permanentemente. Está formada por documentos de diferente naturaleza y formato, desde páginas HTML hasta archivos de imágenes pasando por gran cantidad de formatos estándar y propietarios, no solamente con contenido textual, sino también con contenido multimedia. (Baeza-Yates, 2000).

La Recuperación de Información (RI) no es un área nueva, sino que se viene desarrollando desde finales de la década de 1950. Sin embargo, en la actualidad adquiere un rol más importante debido al valor que tiene la información. Se puede plantear que disponer o no de la información justa en tiempo y forma puede resultar en el éxito o fracaso de una operación. (Baeza-Yates, 2000).

Pero, ¿Qué se entiende concretamente por "Recuperación de Información"? Para Ricardo Baeza-Yates "*la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información*".

Años antes, (Hill, 1983) propuso una definición amplia que plantea que el área de RI "*es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información*".

Cabe aclarar que en las definiciones anteriores los elementos de información son no estructurados, tales como documentos de texto libre o semi-estructurados, como lo son las páginas web.

Croft (W.B, 2000) estima que la recuperación de información es "*el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental*".

Por otro lado, (Korfhage, 1997) definió la RI como "*la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta*".

Ciertamente, es un área amplia, donde se abarcan diferentes tópicos, algunos computacionales como el almacenamiento y la organización; y otros relacionados con el lenguaje y los usuarios como la representación y la recuperación propiamente dicha.

Se puede de cierta manera resumir que la recuperación de la información es la disciplina encargada de mantener una estructura de manera organizada de la información para responder así de una manera eficiente a las necesidades de un usuario.

1.3.3 Extracción de información desde la web.

La extracción de la información es el área de la ciencia y la tecnología que trata de la identificación, clasificación y estructuración en clases semánticas de información específica encontrada en fuentes no estructuradas, para así permitir su posterior tratamiento automático en tareas de procesamiento de la información**.** (R., 1997).

Dada una colección de documentos, tiene como objetivo, identificar y extraer de los mismos aquellos hechos y relaciones relevantes para un dominio particular, ignorando la información extraña e irrelevante.

1. La información obtenida se devuelve de forma estructurada. (R., 1997).
2. Ha de establecerse a priori que constituye un hecho/relación. Relevante. (R., 1997).
3. Sistemas muy especializados de dominio acotado. (R., 1997).

El tratamiento automático de la información facilita a los usuarios la manipulación, evaluación y utilización de grandes cantidades de documentos, tarea que mediante técnicas completamente manuales no se podría realizar. Esta situación se encuentra afectada por la alta disponibilidad de documentos en Internet y la existencia de fuentes de información “en línea”, como por ejemplo las agencias de noticias y periódicos digitales. Una de las técnicas que intenta brindar soluciones a algunos de estos inconvenientes es el resumen automático de textos, proceso por el cual se intenta identificar la información sustancial de un texto para generar una versión abreviada. (Mani, et al., 1999).

Un concepto relacionado con el de resumen de un texto es el de “idea principal”. Para Pardo (T.A.S, 2003) todo texto posee una idea principal y es posible identificar aquella oración que mejor la expresa. Se puede plantear entonces una analogía entre esta oración y un resumen mínimo de un texto, de solo una oración.

La idea principal de un texto es un atributo importante que puede ser útil para diversas aplicaciones, tales como resúmenes automáticos, clasificación de textos y obtención de nuevos metadatos para documentos web. En este último caso, esta información se puede asociar al posicionamiento de los documentos en un sistema de recuperación de información. De manera general, el proceso de extracción de la idea principal puede ser visto como uno de selección de características de un documento.

1.4 Formatos de redifusión de contenidos web.

La redifusión web se ha concretado en el servicio que un sitio web ofrece a usuarios individuales, consistente en mantenerlos permanentemente actualizados sobre sus contenidos, informándoles sobre la renovación de sus titulares y de fragmentos de sus páginas web. A los usuarios receptores de

este servicio se les denomina suscriptores del sitio web original, ya que deben solicitarle de alguna manera dicho servicio.

Las fuentes web suelen codificarse en XML, aunque el formato puede ser cualquier otro que pueda transportarse mediante HTTP, como son HTML o JavaScript. Las dos principales familias de formatos de redifusión web son Atom y RSS. Recientemente el término RSS (Sindicación Realmente Simple) se ha usado indistintamente para referirse también a cualquiera de los formatos de fuentes web, ya sea Atom o RSS.

1.4.1 Formato de redifusión Atom.

Atom es una manera sencilla de leer y escribir información en la web, lo que le permite mantener fácilmente un seguimiento de más de un sitio en menos tiempo y sin problemas para compartir sus palabras e ideas mediante la publicación en la Web. Atom se comenzó a diseñar en 2003 como una alternativa al formato RSS, que había nacido en 1999 en el seno de Netscape, el navegador más popular de esa época incipiente de la Web. En 2005 apareció la versión 1.0 de Atom. Su código se puede reutilizar dentro de otros que usen el lenguaje de programación XML (la base tanto de Atom como de RSS), una idea pensada para favorecer su extensión. Con este formato se pueden agregar diversas fuentes dentro de un contenido y, al mismo tiempo, mantener la información de su creador intacta, con enlaces al sitio original. (WIKIPEDIA, 2013).

Atom es un formato de documento basado que describe las listas de información relacionada conocida como "feeds". Los feeds permiten que los programas busquen actualizaciones del contenido publicado en un sitio Web. Para crear uno, el propietario de un sitio Web puede usar algún software especializado, como un Sistema de gestión de contenido que publica una lista de artículos recientes en un formato estándar legible por máquinas. La fuente web puede ser descargada por sitios web que redifunden el contenido usando dicha fuente, o por un agregador que permiten que los lectores en Internet se suscriban y vean los contenidos de la misma.

Una fuente web puede contener entradas, que pueden ser encabezados, artículos completos, resúmenes y/o enlaces al contenido de un sitio web. Los componentes de este formato organizaron el grupo de trabajo IETF Atom Publishing Format and Protocol. El formato de redifusión Atom fue publicado como un "estándar propuesto" de la IETF en el RFC 4287, y el protocolo de comunicación se publicó como RFC 5023.

Los Atom feeds están compuestos por una serie de ítems conocidos como entradas, cada uno con un conjunto de metadatos adjunto. (Sayre, 2005).

Etiquetas	Descripción
Title	Descripción del título del feed.
Link	URL de la página a la que corresponde el canal.
Updated	Indica la última vez que el feed fue actualizado.
Author	Nombre de un autor del feed.
Id	Identifica al feed utilizando un sistema único y permanente.
Entry	Un documento Atom no requiere un elemento feed para contener un elemento entry. Un elemento entry puede ser parte de un feed y también puede ser su propio documento. Independiente de ello el formato de los subelementos no varía. Un elemento entry es equivalente al elemento analizado en la declaración de los formatos RSS 1.0 y 2.0.
Title	Título de que describe la entrada.
Link	URL de la entrada.

Tabla 1 - Elementos XML que componen Atom 1.0 (Johnson, 2006).

Atom facilita la exportación de los contenidos generados por un medio a otro blog o sitio web. Incluso contempla la posibilidad de importar todos los artículos de un sitio a otro sistema de manejo de contenidos distinto. A partir de él, ha nacido hAtom, un microformato que es el fundamento de los Web Slices de Internet Explorer 8. Se trata de pequeñas ventanas emergentes que aparecen superpuestas en la pantalla del ordenador cuando se publica una novedad en los sitios a los que el usuario se suscribe. Atom estandarizó el código que se debe introducir para que los navegadores reconozcan la existencia de sindicación dentro de un sitio web, para después mostrar el icono correspondiente en la barra de navegación.

La tecnología Atom, que surge como alternativa de RSS, como ya se mencionaba anteriormente, no cuenta con un gran alcance para su puesta en marcha, debido al pobre evolucionar que ha tenido, por tal motivo se decidió valorar y caracterizar el formato RSS, para su posible integración al sistema a desarrollar.

1.4.2 Formato de redifusión RSS.

RSS son las siglas de Really Simple Syndication, un formato XML para indicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos. El formato permite distribuir contenidos sin necesidad de un navegador, utilizando un software diseñado para leer estos contenidos RSS. Fue desarrollado específicamente para todo tipo de sitios que se actualicen con frecuencia y por medio del cual se puede

compartir la información y usarla en otros sitios web o programas. A esto se le conoce como redifusión web o sindicación web.

Al principio, RSS surgió como una iniciativa de una compañía de software llamada UserLand, pero fue utilizada por Netscape para dividir la información de su portal en canales como economía, tecnología, espectáculos, etcétera. A través de los RSS, los usuarios pueden escoger la información que quieran la cual aparezca en su visión personalizada del portal.

RSS es la respuesta práctica para hacer más eficiente la difusión y consumo de información en Internet, debido a que permite promover cualquier sitio y facilita los blogs. La frase “Información es poder” hace recordar el hecho de que, si el usuario se mantiene bien informado, tiene la posibilidad de tomar mejores decisiones y minimizar el riesgo. RSS es un formato estandarizado para compartir encabezados y/o descripciones completas de las notas de periódicos en línea, portales, anuncios clasificados y, sobre todo, blogs o incluso cualquier otra información disponible en un sitio Web y de la que se quieran hacer pruebas para atraer visitantes. Comúnmente, RSS se utiliza como una especie de “¿Qué hay de nuevo?”, pero en realidad, es una forma de sindicación para contenidos publicados en la Web; entendiéndose sindicación como la presencia del contenido de algún medio de comunicación en otros medios del mismo o diferente tipo. (Gonzalez, 2010).

Cuando se habla del término RSS se refiere usualmente a la tecnología completa para distribución de contenidos de los sitios web. Pero un RSS se usa para recoger contenidos publicados en páginas web. Igual que HTML se utiliza para escribir páginas en un formato entendible por los navegadores, para enumerar artículos o páginas dentro de un sitio, en un formato que pueden entender programas denominados lectores RSS o agregadores. En el archivo RSS simplemente están los datos de las novedades del sitio, como el título, fecha de publicación o la descripción.

Para (Yee, 2009) “los feeds son documentos utilizados para brindar información digital actualizada a los usuarios”. La arquitectura de un feed RSS sigue la especificación XML y su vocabulario está compuesto por elementos, subelementos, atributos. (W3C, 2006). Para (Kyrnin, 2008) el modelo básico en cualquier versión RSS tiene las siguientes características:

1- Se trata de un documento basado en XML, por lo consiguiente debe ser un formato bien desarrollado y estructurado (wellformated).

2- El primer elemento de un documento RSS es la etiqueta que indica el comienzo de un archivo RSS.

3- El siguiente elemento es el canal, que contiene los metadatos que lo describen, un título, una breve descripción y la dirección del recurso descrito, normalmente la dirección del sitio o la dirección del feed RSS.

4- Por último, el elemento que especifica cada artículo o contenido del documento RSS. Este elemento al igual que el canal contiene metadatos que lo describen.

Etiquetas	Descripción
Rss versión	Indica la versión de un documento RSS.
Channel	El elemento del canal que contiene metadatos que describen el canal en sí. Debe existir un único elemento channel en el documento.
Title	Descripción del título del feed.
Link	URL de la página a la que corresponde el canal.
Description	Frase o resumen que describe al canal.
Ítem	Representa el contenido del feed. Un canal puede contener n elementos.
Title	Título que describe al elemento.
Link	URL al contenido elemento.
Description	Descripción del elemento.

Tabla 2- Elementos XML que componen la versión RSS 2.0.

El RSS facilita la gestión y publicación de información y noticias webs, que no son más que las noticias digitales publicadas en cualquier sitio web de carácter informativo y noticioso. Es una forma estandarizada de distribución de la información de las páginas web a los lectores de las páginas. Esta información se distribuye a través de las fuentes o canales RSS. Gracias al RSS, los lectores pasan a tener una herramienta útil para mantenerse informado sobre las noticias que le resultan de interés, conservando y almacenando toda la información en un solo lugar que se actualiza de manera automática.

Debido a una amplia investigación, se arriba a la conclusión de que la tecnología RSS está más publicada y más desarrollada que la Atom, siendo esto un factor determinante en el tema de la sindicación de contenidos web, por lo que se decidió utilizar la misma para el desarrollo del sistema propuesto en esta investigación ya que permite al usuario recibir información de distintos sitios y

clasificarla rápidamente, logrando así procesar una mayor cantidad de información en un menor período de tiempo. Esto sería de gran utilidad para la Agencia de Información Nacional a la hora de obtener resultados factibles en el proceso de divulgación de las noticias por los diferentes medios de difusión.

1.5 Herramientas para la recuperación de la información: metabuscadores y buscadores.

En Internet existen numerosas herramientas que se usan para el proceso de recuperación de la información, las mismas brindan una búsqueda de manera rápida y dinámica sobre el tema de interés que se desea indagar. La búsqueda dinámica es “equivalente a la búsqueda secuencial en textos”. Dentro de los SRI en la web, basados en la recuperación de información por medio de palabras clave, se identifican algunas herramientas para el desarrollo de dicha función: metabuscadores y motores de búsquedas o buscadores. (Chang, 2001).

1.5.1 Metabuscadores.

El acceso online a la información puede resultar complejo cuando se trata de acceder a información de numerosas fuentes, más aún cuando cada una de ellas realiza un tratamiento diferenciado de dicha información. Uno de los métodos para solucionar este problema es el uso de metabuscadores**.** Un metabuscador es un tipo de motor de búsqueda que permite buscar a la vez en varios recursos de información: catálogos, texto completo, web, sumarios, citas, bases de datos, portales, bancos de imágenes, directorios y otros buscadores. Un metabuscador, es un buscador de buscadores. Los metabuscadores no tienen base de datos propias, sino que buscan automáticamente en las de otros buscadores, por lo que, partiendo de esta independencia, no almacenan información.

Un metabuscador colecciona las respuestas recibidas y las unifica, “la principal ventaja de los metabuscadores es su capacidad de combinar los resultados de muchas fuentes y el hecho de que el usuario pueda acceder a varias fuentes de forma simultánea a través de una simple interfaz de usuario”. Envían la petición del usuario a todos los motores de búsqueda (basados en directorios y crawlers[2]) que tienen registrados y obtienen los resultados que les devuelven.

Sin embargo cada buscador dispone de su propia sintaxis de búsqueda y en el metabuscador no se puede hacer distinción entre las diferentes sintaxis de cada buscador. Por lo tanto, al buscar información muy específica es mejor emplear buscadores de los que conozcamos la sintaxis. Entre la diversidad de los metabuscadores que existen destacan: ixquick, metacrawler, mamma y dogpile.

[2] Este término se refiere al robot que recopila páginas web para el índice de los motores de búsqueda.

Todo metabuscador debe permitir:
Distribución de la búsqueda por áreas temáticas y por uno o varios recursos a la vez.
Búsqueda simple y avanzada. (METABUSCADORES, 2008).

1. Determinar el número de resultados que se desea obtener por página y el número de resultados que se desea obtener por fuente, es decir, por recurso.
2. Selección del número de registros obtenidos en la consulta.
3. La ordenación de los resultados según varios criterios, incluido el recurso.
4. Supresión de los duplicados.
5. Filtrar los resultados para limitar aún más los resultados de la búsqueda.
6. Exportar los resultados, imprimirlos o guardarlos.

Su funcionamiento se centra en el momento en el que el usuario lanza la búsqueda, donde el metabuscador la dirige a sus motores asociados, devolviendo una lista de resultados que se pueden ordenar según la relevancia.

El principio de funcionamiento de un metabuscador es descrito en los siguientes pasos:

1. Acepta la consulta del usuario. El usuario escribe la consulta en el formulario de búsqueda simple o avanzada y realiza su envío.
2. Convierte la consulta en la sintaxis correcta para cada servicio de búsqueda. La expresión de búsqueda es traducida por el algoritmo a la sintaxis apropiada para cada motor de búsqueda o directorio temático que abarca.
3. Remite la consulta en sus múltiples sintaxis a los diversos SRI. Finalizada la conversión transmite las diferentes ecuaciones de búsqueda a sus respectivos servicios de búsqueda.
4. Espera por las respuestas. Espera un tiempo prudente para recoger la totalidad de las respuestas brindadas por las herramientas de recuperación interrogadas.
5. Analiza los resultados, elimina duplicados. Captura los resultados y elimina las direcciones exactas.
6. Combina los resultados. Todos los aciertos son mezclados conformado una única lista de ítems.
7. Elabora el posicionamiento. Ordena la lista por relevancia de los documentos.
8. Entrega los resultados pos procesados al usuario.

Algunos de los metabuscadores efectúan un proceso de clasificación para agrupar información repetida, mientras que otros no llevan a cabo ningún tipo de análisis.

La ventaja principal de los metabuscadores es que amplían de forma notoria el ámbito de las búsquedas que realizamos, proporcionando mayor cantidad de resultados. La forma de combinar los resultados depende del metabuscador empleado. Además permiten la creación de consorcios, el uso de licencias nacionales, agrupaciones de bibliotecas (coaliciones de compras, consorcios virtuales formados por agregadores de contenidos, consorcios especializados) y bibliotecas digitales.

Sin embargo, la mayor parte de los metabuscadores en realidad ofrecen resultados muy poco exhaustivos, carecen de gestión de búsquedas booleanas, no admiten las búsquedas por etiquetas HTML ni otras características avanzadas específicas de los buscadores individuales y en ellos se detecta la ausencia, difícilmente justificable, de reconocidos servicios de consulta. (METABUSCADORES, 2008). Para profundizar en los resultados de la consulta de un metabuscador se debe acudir a los buscadores originales.

El uso de un metabuscador tendrá como salidas negativas:

1. Sobrecargar el servidor del metabuscador (el ordenador donde se encuentra instalado ese servicio), ya que éste ha de contactar con los buscadores para gestionar las consultas.
2. Sobrecargar la red. Los metabuscadores envían la pregunta a los distintos buscadores aunque estén muy distantes del usuario y no utilizan ni los duplicados ni las versiones nacionales de los buscadores.
3. Sobrecargar los servidores de los buscadores, ya que los metabuscadores siempre lanzan las consultas a varios aunque se localicen suficientes referencias relevantes accediendo sólo a uno de ellos. Es probable que el usuario seleccione más servicios de los estrictamente necesarios para garantizar un gran número de resultados.

Estas son desventajas considerablemente notables a la hora de la integración de un metabuscador a cualquier aplicación. No existe como tal un metabuscador específico que permita su interacción a través de servicios webs, sino que estos, buscadores de buscadores, como ya se mencionaba anteriormente, utilizan los servicios web de los motores de búsqueda que se indexan en su contenido. Por lo que sería redundante y poco eficaz, utilizar un metabuscador en un sistema que cuente en su implementación con la asociación de algún motor de búsqueda. Debido a estas razones y a las desventajas presentadas anteriormente se decidió no hacer uso de esta herramienta de recuperación de la información y valorar y caracterizar otra herramienta encargada de la recuperación de la información, en este caso, los motores de búsqueda.

1.5.2 Buscadores o motores de búsqueda.

Los buscadores, o también denominados motores de búsqueda son robustas aplicaciones que manejan también grandes bases de datos de referencias a páginas web recopiladas por medio de un proceso automático, sin intervención humana. Uno o varios agentes de búsqueda recorren la web, a partir de una relación de direcciones inicial y recopilan nuevas direcciones generando una serie de etiquetas que permiten su indexación y almacenamiento en la base de datos. Un motor no cuenta con subcategorías, sino con avanzados algoritmos de búsqueda que analizan las páginas que tienen en su base de datos y proporcionan el resultado más adecuado a una búsqueda. También almacenan direcciones que les son remitidas por los usuarios.

Entre los motores de búsqueda se pueden mencionar Altavista, Lycos, Alltheweb, Hotbot, Overture, Askjeeves, Direct Hit, Google, Microsoft Network, Bing, Yahoo!, entre otros. La presente investigación se centra en la descripción y análisis de los hoy considerados principales motores de búsquedas: Yahoo, Bing y Google, debido a que son fáciles de usar, rápidos, fiables y además los desarrolladores no tendrán que preocuparse en añadir datos estructurados para cada buscador sino que imperará un lenguaje común para todos permitiendo optimizar el indexado de las páginas y, por tanto, mejorar el posicionamiento en las páginas de resultados**.**

1.5.2.1 Yahoo!

Yahoo!, Inc., es una empresa de medios con sede en Estados Unidos, cuya misión es ser el servidor global de Internet más esencial para consumidores y negocios. En la primavera de 1994, Jerry Yang y David Filo eran estudiantes de postgrado de la Universidad de Stanford. Filo había descubierto la existencia del navegador Mosaic (navegador gráfico para visualizar páginas web). Para llevar el registro de todas las páginas que visitaban los fueron organizando por temas y luego lo publicaron en la Web. La llamaron "Vía rápida de Jerry a Mosaic". Este se convirtió en un nombre tan conocido como el de Wal-Mart y sus sustitutos subsiguientes no fueron mucho mejores: "Guía de Jerry Yang a la WWW" y "Guía de Jerry y Dave a la Word Wide Web". Yang y Filo reemplazaron todas ellas con algo más adecuado para su directorio: Yahoo, lo que en inglés significa "una persona ruda y sin refinamiento". Después de escoger "Yahoo", Yang y Filo decidieron añadir el signo de exclamación. El tráfico de Yahoo!, sobrepasó las 100,000 visitas por día para finales de 1994. Yahoo! hizo su aparición pública en el mercado de valores de Nueva York en el índice NASDAQ el 12 de Abril de 1996, vendiendo 2.6 millones de acciones bajo el símbolo YHOO. Yahoo! es la web más visitada en la actualidad. La red mundial de Yahoo! despliega 3400 millones de páginas web diarias, según los datos de Octubre de 2005.

Entre las características que posee Yahoo! se puede mencionar:

1. Permite la búsqueda independiente de imágenes, vídeos, noticias y compras.

2. Ofrece conceptos relacionados de las búsquedas de forma online.
3. Está integrado con sistemas de microblogging [3]como Twitter.
4. Detecta el tipo de búsqueda y ofrece búsquedas sobre ese tema en sitios relacionados.
5. Ofrece búsquedas seguras.

Entre los servicios que Yahoo! ofrece a sus usuarios están Yahoo! Correo, Yahoo! Messenger, Yahoo! Group, Yahoo! Juegos, Yahoo! Compras, Yahoo! Subastas. Organiza todo en carpetas para un mejor control filtro de span (correos no deseados) y virus, posee un lector de RSS, permite atajos con el teclado y además tiene como soporte tutoriales animados que se enlazan con el mismo buscador.

Sin embargo, la publicidad de la web y la línea insertada en los emails que se envían solo funciona a toda su capacidad con los navegadores IE, Firefox y ópera, presentando solamente un corrector ortográfico en el IE y una sola interfaz alternativa, la de Yahoo clásico.

A pesar de ser un potente buscador, en el presente trabajo se descartó cualquier posibilidad de hacer uso de algunos de sus servicios web y de hacer uso del mismo como motor de búsqueda, puesto que, luego de una ardua y profunda investigación, se llegó a la conclusión de que, acorde al objetivo actual que se persigue para la aplicación, su uso es muy pobre, presenta escasa documentación y una muy pobre actualización de sus servicios, dejando depreciados la mayoría de ellos.

1.5.2.2 Bing.

Bing (anteriormente Live Search, Windows Live Search y MSN Search) es un buscador web de Microsoft. Presentado por el director ejecutivo de Microsoft, Steve Ballmer el 28 de mayo de 2009 en la Conferencia *All Things Digital* en San Diego. Fue puesto en línea el 3 de junio de 2009 con una versión preliminar publicada el 1 de junio del 2009. Una de las grandes novedades de este nuevo buscador de Microsoft, es la opción que permite mostrar las páginas Web, pre visualizar vídeos o ampliar el tamaño de las imágenes sin tener que salir de la página del buscador. Algo que sin duda puede ser muy útil para muchos, pero muy molesto para el resto. Ya que podría entorpecer el libre navegar por Internet. Bing presenta diversas características que lo convierten día a día en un potente buscador. Entre ellas podemos encontrar (WIKI, 2010):

Características de la interfaz

- Vínculos. En ciertos resultados de búsqueda, la página de resultados de búsqueda también muestra vínculos de sección dentro del artículo.

[3] Es una especie de mezcla entre chat, foros, blog y el "estado" que se pone en el Messenger.

- Mejorada vista donde se puede ver información del sitio de terceros dentro de Bing.

Características de Media

- Vista previa del vídeo en miniatura donde, por desplazarse sobre una miniatura de vídeo, el vídeo automáticamente inicia la reproducción.
- Búsqueda de imágenes con desplazamiento de imágenes en páginas de resultados que tienen valores ajustables de tamaño, diseño, color, estilo y personas.
- Búsqueda de vídeo con configuración ajustable de longitud, tamaño de pantalla, resolución y de fuente.

Respuestas inmediatas

- El paquete de seguimiento. Cuando un usuario escribe el nombre de la compañía de transporte y el número de seguimiento Bing proporcionará información de seguimiento directo.
- Ortografía. Se cambió mal con frecuencia los términos de búsqueda para la alternativa más comúnmente escrito. Esta función no puede ser deshabilitada o evitarla.

Información local

- Actual información de tráfico
- Listado de empresas
- Listado de gente

Además de buscar páginas Web, Bing también ofrece las siguientes ofertas de búsqueda:

Bing Noticias es un agregador de noticias y ofrece noticias resultados relevantes a la consulta de búsqueda desde una amplia gama de noticias en línea y servicios de información.

Bing Videos permite al usuario buscar rápidamente y ver vídeos en línea de sitios web distintos. La característica de vista previa inteligente permite al usuario ver instantáneamente una corta de vista previa de un vídeo original. Bing Vídeos también permiten a los usuarios tener acceso a contenido editorial de vídeo desde MSN Video.

Búsqueda visual permiten a los usuarios refinar sus consultas de búsqueda para obtener resultados estructurados a través de galerías de imágenes de dato-agrupación similar a grandes catálogos en línea.

Los servicios de Bing incluyen la Bing API y Bing Maps. A su vez, Bing API incluye un acceso simple y multi-protocolo para acceder a fuentes de los tipos Image, InstantAnswer, MobileWeb, News,

Phonebook, RelatedSearch, Spell, Translation, Video y Web. Bing API soporta comunicación basada en XML**,** JSON **y** SOAP, además de incorporar nuevas funcionalidades como tipado fuerte, manejo de errores y nuevos tipos de los que soportaba anteriormente Live Search como RelatedSearch que entrega resultados de interés lateral que incorpora anuncios a una aplicación. Los servicios de Bing no tienen un límite de solicitudes.

La API de búsqueda de Bing permite a los desarrolladores agregar funcionalidad de búsqueda a un sitio web, crear aplicaciones de consumo exclusivo de la empresa, o desarrollar nuevas aplicaciones webs híbridas. La API de búsqueda de Bing ofrece una gran variedad de fuentes (o tipos de resultados de búsqueda). Se puede solicitar un tipo de fuente única o múltiples tipos de código con cada consulta.

A pesar de que Bing, como buscador, es considerado uno de los fuerte motores de búsquedas del mundo debido a sus características y potencialidad de búsqueda, se descarta su uso en el desarrollo del actual sistema debido a medidas políticas tomadas por Microsoft sobre este buscador, dejando inactiva la prestación de sus servicios. Por lo que, debido a lo investigado a profundidad, se decidió analizar y caracterizar los servicios del potente motor de búsqueda de Internet, Google.

1.5.2.3 Google.

Google fue fundado en septiembre de 1998 por dos estudiantes de doctorado de Stanford, su objetivo era conseguir información relevante a partir de una importante cantidad de datos. Crearon un algoritmo para la búsqueda de datos denominado *PageRank*. Esta tecnología se convertiría más tarde en el corazón que haría funcionar a Google.

En Enero de 1996 iniciaron su colaboración en un buscador llamado BackRub sobre el cual más adelante se construiría Google. (Hernandez, 2011).

Google sólo muestra aquellas páginas que incluyen todos los términos de la búsqueda. Para proporcionar resultados más exactos, Google solo busca palabras completas, es decir no utiliza el truncado automático, únicamente devuelve los términos que ingresamos en la caja de búsqueda. Cada resultado de búsqueda Google contiene un fragmento de la página Web que muestran el contexto en el que los términos aparecen en esa página. Esto proporciona una gran visión de búsqueda para el usuario, ya que no restringe la misma por ningún criterio, sino que a la hora de realizar una búsqueda, se puede obtener cualquier tema sin importar la manera en que se haga, proporcionando al usuario como resultado una breve descripción acerca de lo indagado, dándole así noción de lo que realmente le es de interés.

La arquitectura de Google está diseñada para guardar todos los documentos que se encuentren en el rastreo. Esto tiene un propósito secundario de auxiliar a otros investigadores para que puedan llegar rápido al servidor, procesar datos y obtener resultados interesantes que habría sido difícil de recolectar sin la existencia de esta base de datos. El PageRank[4] utiliza una algoritmo de mapeo para poder calcular la prioridad que debería tener una página específica, y así, poder saber cuáles deberían estar entre las primeras diez. La arquitectura del motor de búsqueda de Google se divide a grandes rasgos en cuatro clasificaciones o subprocesos: El rastreo, la indexación, la clasificación y la búsqueda. Sería de gran utilidad contar en la aplicación con esta arquitectura que presenta este motor de búsqueda ya que permitiría una mayor exactitud y rapidez a la hora de la recopilación de información y principalmente de noticias para su posterior procesamiento en la AIN.

Google brinda interacción con diferentes servicios web a través de sus API's. Las mismas son los métodos que Google ofrece a los desarrolladores para que puedan hacer peticiones a Google desde sus propias aplicaciones. Los desarrolladores pueden hacer peticiones a Google mediante el uso de varios lenguajes, como Java, Perl o Visual Studio .NET, entre otros. Las aplicaciones que escriben los desarrolladores se conectan remotamente con el servicio Web API de Google. Esta comunicación se realiza mediante un protocolo llamado SOAP (Simple Object Access Protocol) el cual está basado en XML, y se usa para el intercambio de información entre aplicaciones. Se pueden encontrar diferentes API's que basan su funcionamiento a través de servicios web, entre ellas destacan Google Search API, Google Search Ajax API, Google Soap Search API, entre otras.

En Mayo del año 2002 Google lanzó un servicio denominado "***Google SOAP Search API***" el cual consiste en una API que permite a desarrolladores acceder a sus servicios de búsqueda desde sus propias aplicaciones utilizando los estándares SOAP y WSDL. Por lo tanto, una vez que se conoce la descripción WSDL de las operaciones de Google, el desarrollador podrá generar peticiones SOAP y procesar los mensajes de respuesta recibidos en su propia aplicación.

Google SOAP Search API es un servicio que permite a desarrolladores buscar y manipular información de la web de una manera rápida y sencilla. Los desarrolladores escriben programas que se conectan de forma remota al servicio de Google vía SOAP para el intercambio de información. Las funcionalidades de este servicio son: hacer consultas al índice de Google el cual cuenta con billones de páginas web y recibir los resultados en datos estructurados, acceder a la información de la caché de Google y chequear la correcta escritura de una palabra con las sugerencias ofrecidas por Google. Algo importante para destacar es que el servicio utiliza la misma sintaxis de búsqueda que el sitio Google.com, por lo tanto, los usuarios del servicio podrán realizar las búsquedas de la misma forma

[4] El PageRank es la referencia que utiliza Google para indexar y clasificar los contenidos de las páginas web, de acuerdo a la importancia que tiene cada una en la comunidad.

en que lo harían en el sitio de Google. Poder interactuar con un servicio web mediante SOAP sería una forma para el desarrollo de esta investigación; SOAP es un protocolo independiente del lenguaje, multiplataforma, muy simple de usar, se adapta a las normas o estándares de la entidad y es muy sólido en cuanto a las funciones que describen los XML, siendo esto un potencial a la hora de combinarlo con RSS. Sin embargo, esta API quedó depreciada por Google a finales del 2006 reemplazando dicha API por Ajax Search API, por tal motivo esta API fue descartada para su utilización en el presente proyecto a desarrollar.

Google Search AJAX API es una API creada por Google que permite a los desarrolladores web acceder a los servicios de búsqueda de Google en sus propias páginas web. Se trata específicamente de una biblioteca JavaScript que proporciona unos objetos simples que realizan búsquedas en línea sobre un número de servicios de Google; todo ello a través de programación JavaScript y XML (AJAX).

Se ofrece gratuitamente para que los desarrolladores puedan probarla, ya que aún no existe ninguna versión comercial. Pero para poder empezar a utilizarla es necesario registrarse previamente y obtener una clave que otorga un máximo de 1000 peticiones por día. Esto a fines de evitar que otro servidor Web funcione paralelamente a Google. Serviría de mucho contar con el servicio de esta API en el presente sistema.

Constituye una de las tecnologías que componen los servicios web. Esto significa que el usuario puede acceder a un sitio web que implementa este API, realizar una consulta y obtener los resultados de la misma en el mismo sitio web sin saber que se está empleando un servicio Google.

La API de búsqueda AJAX le permite integrar fácilmente a los desarrolladores diversos mecanismos de búsqueda de Google muy poderosos basados en "controles" de una página Web con la codificación relativamente mínimos. Además de permitir integrar la búsqueda de noticias o la búsqueda de algún blog en cualquier página web. Algunas de las características que poseen estos controles son:

1. Buscar en la Web: Se trata de un campo de búsqueda de entrada tradicional, donde, cuando una consulta se introduce una serie de resultados de búsqueda de texto aparecen en la página. Esta característica da un amplio margen de búsqueda sobre un factor determinado, sería una importante característica a incorporar en esta propuesta para la AIN, permite a los redactores de la noticia tener un amplio sentido de búsqueda sobre cierto tema y criterio de interés siendo fácil así el proceso de búsqueda de una noticia determinada de un sitio determinado para su posterior procesamiento.

2. Búsqueda local: Con la búsqueda local, un mapa de Google se sobrepone con un campo de búsqueda y los resultados de la búsqueda se basan en una ubicación específica. Con la búsqueda local los redactores de la noticia de la AIN pudieran realizar búsquedas específicas de qué y dónde es lo que quieren buscar, sin tener que obtener un gran número de resultados de búsqueda para entonces seleccionar cual sería de su interés, minimizando así el tiempo de recopilación de la información.
3. Buscar Medias: La Búsqueda de medias de AJAX proporciona la capacidad para ofrecer búsqueda de una media convincente junto con los resultados que se acompañan de búsqueda de vídeo. A la hora de confeccionar la noticia en el proceso llevado a cabo por los redactores de la AIN, es necesario contar con alguna media, dígase audio, imagen o vídeo, para su posterior procesamiento y divulgación; contando con esta opción en la presente aplicación a desarrollar para la Agencia de Información Nacional, los redactores de la misma no tienen que hacer ningún proceso manual de obtención de una media, sino que de acuerdo a la noticia que busque, al mismo tiempo podrá tener su media asociada teniendo en cuenta la sección temática a la que pertenece la noticia, ganando en tiempo y haciendo más eficaz el cumplimiento de su función.

Estas características hacen de Google un fuerte y potente motor de búsqueda, pero sin embargo se guiará la integración de dicho motor con el presente sistema hacia la principal característica que identifica el mismo, la Búsqueda de Noticias.

Pese a las restricciones de esta API se puede cargar en la aplicación una biblioteca jquery mediante la API de Java Script de Google y a partir de ahí, se puede utilizar el sistema de búsqueda indexada de Google. Esto sería un gran avance tecnológico en el sistema propuesto para minimizar el proceso de recuperación de la información en la AIN.

1.6 Conclusiones parciales.

Luego de dejar plasmada la situación problemática que dio origen a la investigación y del estudio de los principales conceptos asociados al tema, se definió el uso del formato de redifusión RSS como lector de fuentes web ya que tiene mayor uso y documentación en el ámbito de la redifusión de contenido web. Además se ha realizado un análisis crítico y valorativo acerca de algunas características presentadas por los motores de búsqueda, concluyendo que Google es el candidato idóneo para su incorporación al sistema propuesto. A partir de las dos vías de recuperación de la información seleccionadas se obtendrá una mayor eficiencia en los procesos de redacción que se llevan a cabo en la AIN.

Capítulo 2: Análisis y presentación de la propuesta de solución.

2.1 Introducción.

En este capítulo se describe la propuesta de solución para los problemas de recuperación de la información que existen en la AIN. Se explican las razones del uso de las herramientas para guiar y dar soporte a la solución que sea propuesta. Además con la construcción del modelo de dominio se llega a un acercamiento mayor del problema a resolver. Mientras que el levantamiento de los requisitos de software va a permitir conocer las condiciones y cualidades deseadas por los usuarios.

2.2 Caracterización de las tecnologías, lenguajes y herramientas a emplear.

A continuación se presentan las características de las diferentes tecnologías, lenguajes y herramientas, así como la metodología a utilizar para dar solución al problema científico planteado.

2.2.1 Marco de trabajo.

En el desarrollo de software, un marco de trabajo, o *framework* en inglés, es una estructura de soporte definida en la cual otro proyecto de software puede ser organizado y desarrollado. Típicamente, puede incluir soporte de programas, bibliotecas y un lenguaje interpretado entre otros software para ayudar a desarrollar y unir los diferentes componentes de un proyecto.

2.2.1.1 JQuery.

JQuery es considerado una biblioteca de Javascript, o ambiente de desarrollo. Es un conjunto de utilidades las cuales no necesitan ser programadas y se pueden utilizar de una manera muy simplificada. Brinda la posibilidad de trabajar con AJAX[5], sin preocuparnos de los detalles complejos de la programación. Es software libre y de código abierto, posee un doble licenciamiento bajo la licencia MIT y de la GNU General PublicLicense, Versión 2. Utilizar esta biblioteca en el desarrollo del presente sistema trajo consigo la obtención de una interfaz amigable y presentable para el usuario.

2.2.2 Herramienta de modelado de software.

5 AJAX : Asynchronous JavaScript And XML

La Ingeniería de Software Asistida por Computadora (CASE, por sus siglas en inglés) es un tipo de ingeniería de software en la que se intenta aumentar la eficacia de sus procesos, al soportar la realización de las tareas con el uso de tecnologías.

2.2.2.1 Visual Paradigm.

Visual Paradigm permite modelar todos los artefactos que se obtienen a partir del análisis del negocio y el sistema. Es una herramienta multiplataforma que soporta completamente el ciclo de desarrollo de un software: análisis y diseño, construcción, pruebas y despliegue. Posibilita el modelado de base de datos, requisitos, proceso de negocio, permite realizar todo tipo de diagramas de clases, ingeniería inversa, generar código desde diagramas y generar documentación.

Las razones de su elección sustentan la necesidad de realizar la ingeniería inversa, desde el sistema gestor de bases de datos a los diagramas de entidad-relación para obtener el modelo actual e incorporarle los nuevos cambios y a su vez generar de forma automática la nueva base de datos del sistema.

2.2.3 Lenguaje del lado del servidor.

PHP es un lenguaje de programación interpretado, diseñado originalmente para la creación de páginas Web dinámicas. La mayoría de su sintaxis está prestada de los lenguajes de programación C, Java y Perl, con la inclusión de algunos rasgos únicos de PHP. La meta del lenguaje es permitir a los que desarrollan sitios Web escribir rápidamente páginas generadas dinámicamente.

El uso de PHP para el sistema propuesto tiene como base principal las ventajas de este lenguaje, pues admite la programación orientada a objetos permitiendo programar el software de forma que esté organizado en la misma manera que el problema que se trata de modelar. Existe un amplio conocimiento del mismo y contiene una amplia documentación situada en su página oficial y está respaldado por una activa y numerosa comunidad de desarrolladores.

2.2.4 Lenguajes del lado del cliente.

Lenguaje de Hipertexto Marcado (HTML por sus siglas en inglés): Es un lenguaje sencillo de estructura jerárquica muy usado en las páginas web, definido a base de etiquetas que permiten crear hipertextos. Es usado en el sistema a desarrollar para describir la estructura y el contenido en forma de texto, así como para complementar el texto con objetos tales como imágenes. Permite crear hiperenlaces haciendo posible la relación con otras fuentes de información. (Hernandez, 2011).

JavaScript: Es un lenguaje de script multiplataforma orientado a objetos. Debido a que es un lenguaje pequeño y ligero se decidió hacer uso del mismo para la presente aplicación; además no es útil como

un lenguaje independiente, más bien está diseñado para una fácil incrustación en otros productos y aplicaciones, tales como los navegadores web. (Eguiluz, 2009).

CSS: Es un lenguaje para controlar los estilos de presentación de documentos definidos con HTML y XHTML, separando los contenidos de su presentación. (Eguiluz, 2009). CSS se utilizó para definir el aspecto del contenido, es decir, su color, tipo de fuente y tamaño de letra de los párrafos, el espacio entre párrafos, la tabulación con la que se muestran los elementos de una lista.

2.2.5 Metodología de desarrollo de software.

Las metodologías de desarrollo de software son un conjunto de procedimientos, técnicas, herramientas y un soporte documental que ayudan y guían el desarrollo de un producto de software. Se refiere a un *framework* (marco de trabajo) que es utilizado para estructurar, planear y controlar el proceso de desarrollo de sistemas informáticos, además de definir el conjunto de actividades que guían los esfuerzos de las personas implicadas en el proyecto. Se pueden clasificar en dos grupos generales: las pesadas (o tradicionales) y las ligeras (o ágiles).

Las metodologías tradicionales están orientadas al control de los procesos, estableciendo rigurosamente las actividades a desarrollar, herramientas a utilizar y requieren de una documentación considerable. Se ajustan más a proyectos grandes, con mayor alcance de tiempo. Por otra parte las metodologías ágiles están orientadas a la interacción con el cliente y el desarrollo incremental del software, para ello se muestran versiones parcialmente funcionales del software en intervalos cortos de tiempo, para que el cliente pueda evaluar y sugerir cambios en el producto según se va desarrollando.

2.2.5.1 Proceso Unificado de Software.

Entre las metodologías tradicionales más utilizadas está el Proceso Unificado de Software (*RUP* por sus siglas en inglés). Entre sus principales características se encuentra que es dirigido por casos de uso, lo que define una guía en los procesos de diseño, implementación y prueba, que responde directamente a los requisitos funcionales del software. Otra característica importante es la de ser centrada en la arquitectura, que permite tener una perspectiva clara del sistema completo, además de ser iterativo e incremental, definiendo que por cada iteración se creen incrementos del producto final.

Una de las formas de lograr un registro detallado de la solución y tener una perspectiva amplia de qué es lo que se quiere desarrollar es la utilización de la metodología de desarrollo de software RUP, la que será utilizada para documentar los procesos sustanciales en el desarrollo del diseño del sistema generando una gran cantidad de artefactos, lo cual representa una garantía para la

continuidad del trabajo del proyecto y la futura implementación de las funcionalidades modeladas. Es una de las metodologías más utilizada para el análisis, implementación y documentación de sistemas orientados a objetos.

2.2.6 Sistema Gestor de Base de Datos PostgresSQL.

Es un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el sistema de gestión de bases de datos de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales. PostgresSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando. (PostgreSql, 2012).

PostgresSQL constituye un excelente motor de bases de datos, es un sistema fácil de administrar y robusto. Posee la facilidad de ser multiplataforma, o sea, que se podrá usar en cualquier sistema operativo, estas características han permitido que se haya escogido su uso para la realización del sistema recuperador de noticias publicadas en Internet para la AIN.

2.3 Modelo de Dominio.

Utilizando la notación UML, un modelo del dominio se representa como un conjunto de diagramas de clases en los que no se define ninguna operación. Y pueden mostrar objetos del dominio o clases conceptuales, así como asociaciones y atributos. Su objetivo es lograr la representación de las clases conceptuales del mundo real significativas en un dominio del problema, no de componentes de software, aunque algunos objetos del modelo pueden terminar siéndolo. (Chang, 2001).

Este tipo de modelo puede utilizarse para capturar y expresar el entendimiento obtenido en un área cualquiera. Son similares a los mapas conceptuales utilizados en el aprendizaje y son utilizados por el analista como un medio para comprender el negocio a informatizar por el sistema.

2.3.1 Conceptos y principales eventos del entorno.

El **Redactor** interactúa con el **Winkertin**, herramienta encargada de la recuperación de la información en **Internet**, la cual obtiene la **Noticia en Formato Texto**. Por otro lado, interactúa con los **Motores de Búsquedas** para consultar las fuentes de información, entre las cuales se encuentran las **Fuentes Web** existentes en el mundo entero, con el objetivo de obtener **Información Noticiosa** en formato texto para su posterior procesamiento, además de obtener las medias que van a estar asociadas al contenido de la noticia y a su posterior confección. Dicha noticia, conjuntamente con su formato y con las medias obtenidas, constituyen los componentes esenciales para la

redacción de la **Noticia Completa**, con todos sus elementos asociados, lista para su difusión mediante los diferentes medios informativos existentes en la AIN que divulgan información de cualquier categoría, en este caso una **Radio** y un **Canal de televisión** y los medios que divulgan información propia de la entidad, una **Página Web**.

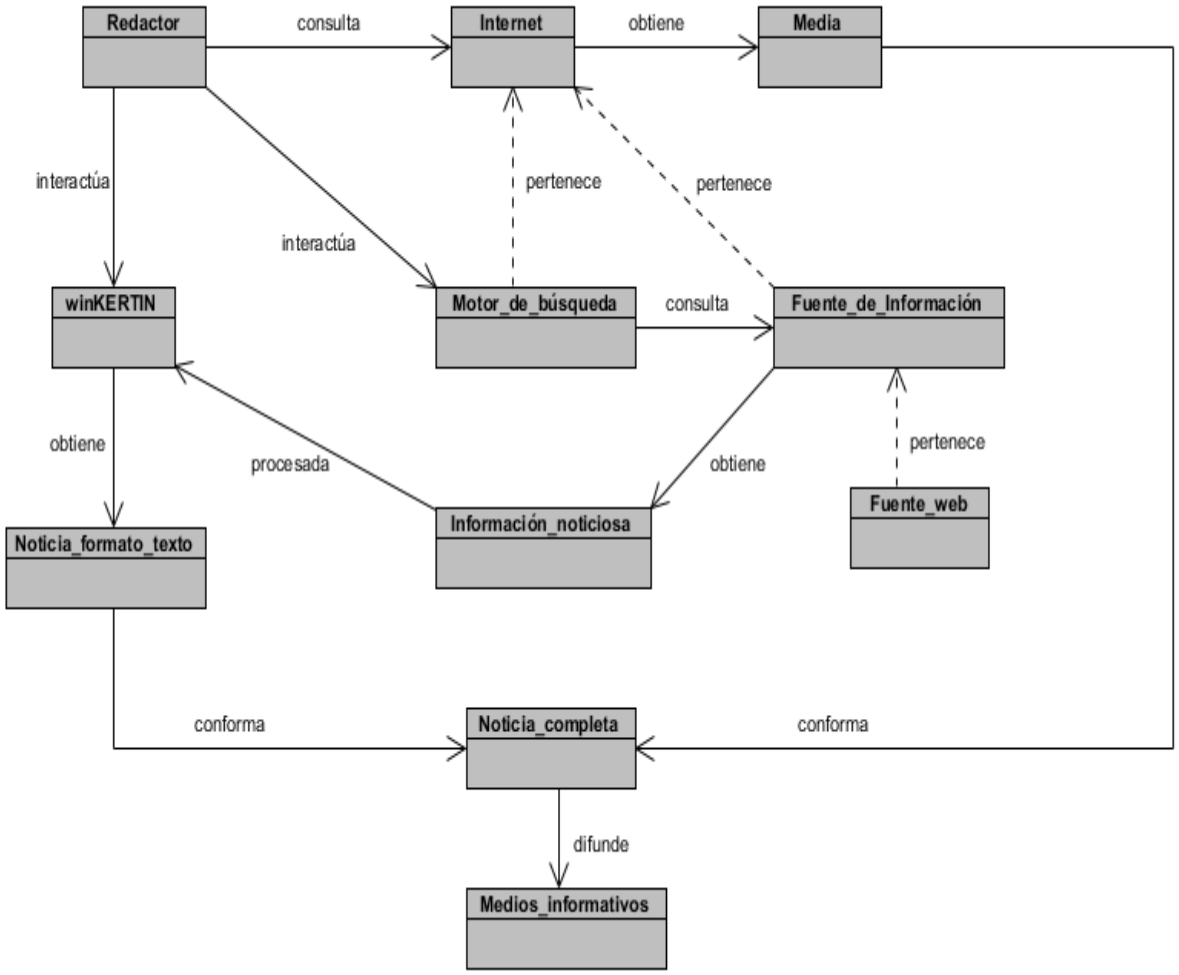


Figura 2: Diagrama Modelo de Dominio.

2.3.2. Descripción de las clases del dominio.

La siguiente descripción de clases ayudará a una mejor compresión de los conceptos presentes en el entorno del problema.

Internet es la red de redes que permite la conexión entre varias computadoras a través de un protocolo de comunicación.

El **Redactor** es el encargado de llevar a cabo el proceso de redacción de la noticia y su posterior divulgación.

Un **motor de búsqueda** es un sistema informático que buscan archivos almacenados en servidores web a través de la indexación.

WinKertin es la herramienta utilizada para la recuperación de información.

La **noticia en formato texto** es el conjunto de información que llega a las personas de forma inmediata y actualizada solamente a través de texto.

La **noticia completa** es el conjunto de información que llega a las personas de forma inmediata y actualizada ya sea a través de texto, imagen, sonido, video o la integración de estos.

Fuente web es un sitio web que a través del formato RSS u otros formatos pueden distribuir sus contenidos y/o ser leídos mediante un lector de fuentes, sin necesidad de acceder al sitio web.

La **radio** es el medio de difusión masivo que llega al radio-escucha de forma personal siendo el medio de mayor alcance ya que llega a todas las clases sociales.

El **canal de televisión** es un tipo de estación emisora que transmite audio y vídeo a receptores de televisión en un área concreta. Hacen sus transmisiones enviando señales de radio especialmente codificadas por el aire, llamada televisión terrestre.

La **página web** es la información electrónica adaptada para la World Wide Web y que puede ser accedida mediante un navegador permitiendo la interacción digital con el usuario.

2.4 Especificación de los requisitos de software.

Según la IEEE Standard Glossary of Software Engineering Terminology 610, define un requisito como:

“Condición o capacidad necesaria para que un usuario resuelva su problema o alcanzar un objetivo que debe cumplirse o poseído por un sistema o componente del sistema para satisfacer un contrato, norma, especificación o formalmente impuesto en un documento”. (Veluchamy, 2010).

Los requisitos son los que permiten guiar el desarrollo de cualquier software hacia un sistema correcto. A través de ellos se definen los objetivos generales concretos de forma tal que tanto el negocio como sus actores se beneficien; se obtiene una descripción correcta de lo que debe hacer el sistema y delimita su alcance; se aumenta la comunicación entre un cliente y el grupo de desarrollo; y sirven de base para la validación, verificación y procedimientos de aceptación de cualquier producto de software. A continuación se mencionan a través de requisitos funcionales y no funcionales, las acciones que el sistema debe ser capaz de realizar.

2.4.1 Requisitos Funcionales.

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir. (Fuentes, 2007). Durante el estudio de los procesos actuales en la redifusión de contenido web, así como el análisis de algunos de los motores de búsquedas más importantes en la actualidad, se determinaron los requisitos con que debía contar el Sistema Recuperador de Noticias Digitales para mejorar el proceso actual de difusión de noticia de la AIN. A continuación se relacionan dichos requisitos:

RF1 Dar la posibilidad de realizar la configuración del proxy a la hora de consultar información en Internet.

RF2 Dar la posibilidad que se puedan consultar diversas fuentes web.

RF3 Permitir la gestión de fuentes web consultadas utilizadas para la extracción de información.

RF3.1 Adicionar fuentes web: Permitir adicionar nuevas fuentes web al sistema.

RF3.2 Eliminar fuentes web: Permitir eliminar fuentes web existentes en el sistema.

RF4 Obtener noticias. El sistema debe ser capaz de obtener las noticias de las fuentes web disponibles.

RF5 Permitir la búsqueda de información de carácter noticioso en Internet mediante la integración de un motor de búsqueda.

RF6 Mostrar las noticias obtenidas. Mostrar al usuario, las noticias obtenidas mediante el motor de búsqueda.

RF7 Permitir la extracción del contenido completo una noticia obtenida a través de una fuente web.

RF8 Permitir la extracción del contenido completo una noticia obtenida a través del motor de búsqueda.

RF9 Almacenar la noticia completa una vez realizada la extracción de sus datos.

RF10 Brindar la posibilidad de guardar directamente la noticia completa, sin tener que acceder a su contenido necesariamente.

RF11 Permitir, al guardar la noticia directamente, su identificación en la interfaz que signifique que ya ha sido guardada.

2.4.2 Requisitos No Funcionales.

Los requisitos no funcionales son propiedades o cualidades que el producto debe de tener. Son características que hacen al producto, atractivo, usable, rápido y confiable. Las propiedades no funcionales, como cuán usable, seguro y agradable, pueden marcar la diferencia entre un producto bien aceptado y uno con poca aceptación. Los requisitos no funcionales se clasifican en múltiples categorías, a continuación se mencionan aquellos necesarios para que el Sistema Recuperador de Noticias Digitales sea integrado de forma exitosa a la AIN.

Requisitos de apariencia o interfaz externa: La apariencia del sistema debe contar con una interfaz amigable, intuitiva, interactiva y simple de usar.

Requisitos de usabilidad: El sistema de forma general debe brindar gran facilidad de uso para personas con poca experiencia con las computadoras, pero con nivel calificado. Para trabajar con el Sistema Recuperador de Noticias Digitales se requiere de conocimientos mínimos en informática, televisión y uso de la web. Las funcionalidades deben ser claras y se debe mostrar la información de forma lógica y correctamente estructurada. El nuevo sistema debe permitir una mejor configuración adaptándolo a las necesidades de la entidad donde se implante, y el cual reduzca los costos en tiempo y complejidad para estas operaciones con relación a la actualidad. Se requiere que pueda ser usado por personas que hablen diferentes idiomas.

Requisitos de portabilidad: El sistema estará realizado para funcionar con el sistema operativo Windows y Linux.

Requisitos de Software: Se debe utilizar Apache en su versión 2.2 como servidor web y PostgresSQL 8.4 o superior como sistema gestor de base de datos. El Sistema Recuperador de Noticias Digitales debe ser accedido a través del navegador Google Chrome en su versión 17 y Mozilla Firefox en su versión 19.5. En caso de que no se tenga un IP real en la estación servidora se deberá hacer uso un servicio proxy para acceder a la interacción con Internet.

Requisitos de Hardware:

1. Procesador: Dual-Core 2.33 GHz.
2. Memoria RAM: 1Gb.
3. Disco Duro: 300Gb.
4. Tarjeta de Red: Ethernet Gigabyte.

Restricciones en el diseño y la implementación: Para la modelación del sistema se utilizará el lenguaje UML mediante la herramienta Visual Paradigm. Se requiere el uso de la arquitectura

Modelo-Vista-Controlador y el uso del lenguaje PHP. La arquitectura debe soportar migrar la interfaz de usuario de forma rápida, para lograr visualizar cualquiera de los cambios que se produzcan.

2.5 Descripción de la solución propuesta.

Después de haber realizado el levantamiento de requisitos del sistema y tener en cuenta varias restricciones, se propone el desarrollo de un sistema que gestione todos los procesos de recuperación de la información noticiosa de la AIN. Dicho sistema dará solución al problema de redacción actual en la agencia, automatizando así un conjunto de funcionalidades que hasta el momento se vienen realizando manualmente.

2.5.1 Recuperación de la información en el sistema.

El Sistema Recuperador de Noticias Digitales presenta dos vías para lograr la tarea de recuperación de la información. La primera, y ya definida previamente, es la utilización de un lector RSS, debido a que hoy en día, con el desarrollo avanzado de la era digital, la gran mayoría de los sitios web en Internet cuentan en su diseño con la integración de un canal RSS, para lograr así mantener actualizado al usuario en el mundo noticioso de una manera eficiente y rápida. A través de la dirección de un sitio web (URL), se realiza el proceso de recuperación de las noticias que están publicadas en dicha dirección, obteniendo un listado con los titulares de las mismas. Antes el usuario deberá configurar el proxy si la dirección a la que está tratando de acceder está regida por esa directiva, en caso contrario se procede normal. La segunda vía, y viéndola como una alternativa a la primera, pero no dejando de ser menos importante, es a través de la integración de un motor de búsqueda, en este caso y definido con anterioridad, Google con su servicio web de Google Search Ajax API, mediante el cual el usuario realiza una búsqueda determinada acerca de un tema de interés determinado, y el sistema busca por medio de una URL, todas las noticias que se relacionen con la consulta de búsqueda hecha por el usuario mostrando finalmente un listado con los titulares de dichas noticias y completando así el proceso de recuperación. Hasta ahora se ha visto cómo se realiza la recuperación de la información en el presente sistema, teniendo como factor común para efectuar dicho proceso, el acceso a la información mediante una dirección URL y dejando las bases listas para la extracción de la información. Sin embargo, el contar solamente con la URL de una noticia no es suficiente para lograr cumplir con éxito el proceso de extracción.

2.5.2 Extracción de la información en el sistema.

El formato de redifusión RSS está compuesto por el formato XML, vista su definición en epígrafes anteriores, el cual brinda una estructura organizada que favorece a la hora de acceder a la información contenida en el mismo. Con la URL obtenida anteriormente en el proceso de recuperación, se procede a parsear el archivo HTML (eliminar etiquetas no deseadas) el cual

pertenece a dicha URL para obtener algo en específico. En este caso solo se quiere obtener del XML el *título de la noticia* y la *fecha de publicación*, esto es posible gracias a su estructura la cual brinda una serie de elementos (ítems) mediante los cuales se puede acceder a los elementos mencionados anteriormente. Cada elemento va a contener un título (title) y una fecha de publicación (pubDate). Para el caso del motor de búsqueda se hace un proceso de manera similar. Mediante la URL se obtienen una serie de resultados r, cada r va a contener una fecha de publicación y un título. Una vez extraídos el título y la fecha de publicación y teniendo constancia de ellos, se procede a extraer los demás datos definidos en la estructura de la noticia. Para ambos casos, el lector RSS y el motor de búsqueda, se opera de manera similar siempre teniendo presente la URL; para obtener *la entradilla o subtitular*, primeramente se parsea el HTML al cual pertenece dicha URL mediante la función *str_get_html,* perteneciente a la clase *simple_html_dom* de php conjuntamente con el modelo de objetos (DOM), una vez parseado el HTML, acto seguido se busca por etiquetas dentro del contenido HTML donde coincida con la etiqueta *sub-título,* en caso de no encontrarse ninguna coincidencia el sistema devolverá como valor para ese elemento: *"Desconocido",* y así para todos los datos que se quieran extraer pero cuyo resultado es nulo o vacío. Para el *resumen* el proceso es el mismo, lo único que cambia es la etiqueta por cual buscar que para este caso será buscar dentro de las etiquetas *meta* y luego buscar dentro de las etiquetas *content* contenidas dentro de las *meta* y seguidamente y por ultimo buscar el resumen o descripción dentro de la etiqueta *description* contenida dentro de las dos mencionadas anteriormente. A la hora de recuperar el *autor* se busca simplemente por la etiqueta *author,* sin embargo como ya se mencionó en epígrafes anteriores, el éxito de la recuperación y extracción depende de cómo este estructurado el HTML. Para obtener el cuerpo de la noticia simplemente se busca por todas las etiquetas *p* y se devuelve el resultado. Para obtener las *imágenes* asociadas a la noticia se realiza un poco de manera diferente, primeramente se buscan todas las imágenes mediante la etiqueta *img,* una vez obtenidas todas las imágenes esto no se hace suficiente para lograr el objetivo trazado, puesto que solo se quiere contar con las imágenes que solo reflejen lo relacionado con el tema de la noticia, para esto se procede entonces a guardar todas las imágenes obtenidas anteriormente en un archivo temporal en el servidor local, para luego, según un patrón definido, extraer la(s) imagen(es) correspondiente al contenido de la noticia, para extraer el audio y el video también es válida la búsqueda por etiquetas, en este caso la etiqueta *audio* y *video,* respectivamente, pertenecientes a HTML5, una vez extraídos ambos elementos se muestran en la interfaz al usuario para que pueda realizar su visualización o escucha y se guardan en el servidor local poseído.

2.5.3 Almacenamiento de la información en el sistema.

Una vez extraída toda la información referente a la noticia por medio de las dos vías mencionadas anteriormente se procede a almacenar la noticia en la estructura definida para esta función, en este

caso, una base de datos operacional. La misma contará con 5 tablas para el almacenamiento de los datos, una primera tabla para almacenar todos los datos referentes a las noticias, una segunda tabla para guardar todo lo referente a las fuentes web agregadas por el usuario al sistema y otras tres tablas las cuales dependen de la tabla *noticia* para almacenar las medias. En este caso, a la hora de almacenar las medias se guardará la dirección local de la media una vez haya sido descargada, como se mencionó en el proceso anterior, esta dirección indicará a que noticia pertenece la media almacenada.

2.6 Modelo de Casos de Uso.

2.6.1 Actores propuesto para el sistema.

Encontrar los actores es uno de los primeros pasos en la definición del uso del sistema. Cada tipo de fenómeno externo con el que debe interactuar el sistema está representado por un actor. Se debe definir cada actor escribiendo una breve descripción que incluya el área de responsabilidad del actor y para qué necesita el actor el sistema. Como los actores representan elementos externos al sistema, no es necesario describirlos en detalle. (Copr., 2006).

Los actores propuestos para el nuevo funcionamiento de la plataforma son los que aparecen en la **Tabla 3**. Su elección está determinada por la aparición del Subsistema de Configuración el cual permitirá la gestión de los roles de usuarios. Por lo que se representa de forma genérica el actor Usuario debido a que el rol que desempeñe en el sistema estará en dependencia de las necesidades de cada cliente. A continuación se proponen los actores y la función que estos desempeñarán.

Tabla 3: Propuesta de Actores del Sistema.

Actor	Descripción
Redactor	Persona responsable de la redacción y divulgación de las noticias a través de los diferentes medios existentes en la AIN.

2.6.2 Casos de Uso del Sistema Recuperador de Noticias Digitales.

La mejor forma de encontrar casos de uso es considerar qué necesita cada actor del sistema, es importante tener presente que el sistema sólo existe para los usuarios y que por lo tanto debe basarse en las necesidades de éstos. Se debe reconocer muchas de las necesidades de los actores a través de los requisitos funcionales realizados sobre el sistema. (Copr., 2006).

Los casos de uso son el componente clave del modelado. Permiten ilustrar las funcionalidades de un sistema y la relación con el actor para cumplir un objetivo. A continuación se expone el diagrama de casos de uso que conforman el Sistema Recuperador de Noticias Digitales.

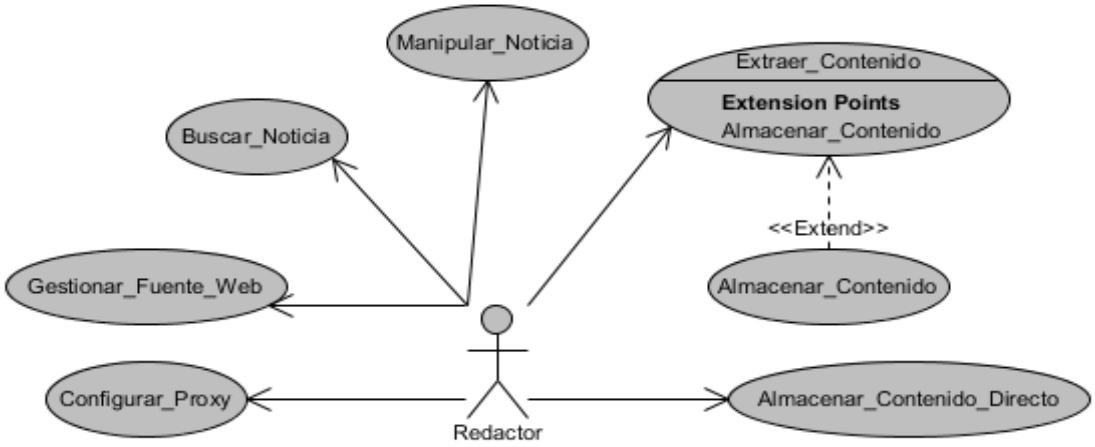


Figura 3: Diagrama de Caso de Usos del Sistema.

El detallar un caso de uso es una de las tareas del flujo de trabajo de Requisitos en RUP y tiene por objetivos, describir uno o varios de los flujos de sucesos del caso de uso con el detalle suficiente para que el desarrollo de software pueda empezar en él. (Copr., 2006).

La plantilla de especificación de caso de uso permite que a través de la descripción textual se pueda detallar cada uno de los aspectos a tener en cuenta en la realización de un caso de uso. Entre esos aspectos se pueden encontrar la descripción del flujo de eventos normal y alterno, la interacción entre actores y el sistema, pre-condiciones y pos-condiciones que deben cumplirse para la realización del mismo, así como el prototipo de interfaz de usuario para cada uno de los escenarios del caso de uso en cuestión.

Para lograr una mayor comprensión de los casos de uso resulta necesario el estudio de la especificación de cada uno de ellos. Para consultar la descripción de los principales casos de uso del sistema se propone visitar el **Anexo 1**.

2.7 Conclusiones parciales.

En este capítulo se seleccionaron las herramientas y tecnologías a emplear en el desarrollo del producto con el objetivo de lograr un software de calidad y altamente funcional. Se analizaron y valoraron todos los requisitos de software planteados por los usuarios finales del sistema para tener así un funcionamiento acorde a la petición del cliente. Como resultado principal de este capítulo quedó plasmada la propuesta de solución para el problema actual del sistema, la cual deja trazado el camino para un mejor entendimiento a la hora de interactuar con las funcionalidades prácticas del software.

Capítulo 3: Construcción de la solución propuesta.

3.1 Introducción.

El presente capítulo describe la arquitectura del sistema y los patrones arquitectónicos a utilizar en el diseño e implementación de la solución propuesta, modelándose la misma a través del diagrama de clases del diseño, además de reflejar como quedó concebido el despliegue de la aplicación. Además de conocer los parámetros para de evaluación para que el presente sistema resulte factible y efectivo.

3.2 Descripción de la arquitectura.

En un sistema informático la arquitectura es el conjunto de decisiones significativas sobre la organización de un software. Para describir y diferenciar una arquitectura se hace necesario emplear estilos arquitectónicos y de diseño que ayuden a definir una estructura para todos los componentes del sistema.

3.2.1 Patrón arquitectónico.

Los patrones arquitectónicos definen la estructura general del software, indican las relaciones entre los subsistemas y los componentes del software y definen las reglas para especificar las relaciones entre los elementos (Clases, paquetes, subsistemas) de la arquitectura. (Roger, 2005). Para describir y diferenciar una arquitectura se hace necesario emplear estilos arquitectónicos que ayuden a definir una estructura para todos los componentes del sistema.

Un estilo arquitectónico es una trasformación impuesta al diseño de todo un sistema. Algunos de estos estilos arquitectónicos son: Arquitectura centrada en datos, Arquitectura de flujo de trabajo, Arquitectura de llamada y retorno, Arquitectura orientada a objetos. (Roger, 2005).

En el sistema se hizo uso del estilo arquitectónico Llamada-Retorno, utilizando el patrón arquitectónico Modelo Vista Controlador (MVC) tomando lo mejor de la arquitectura MVC, garantizando un modelo, una vista y una controladora de forma independiente. El mismo accede a realizar modificaciones en cada una de las partes sin afectar las demás, permitiendo que el desarrollo de aplicaciones sea rápido y sencillo.

Modelo-Vista-Controlador

Modelo Vista Controlador es un patrón de arquitectura de software que separa los datos de una aplicación, la interfaz de usuario y la lógica de control en tres componentes distintos. El patrón MVC se ve frecuentemente en aplicaciones web, donde la vista es la página HTML y el código que provee de

datos dinámicos a la página. El modelo está compuesto por el Sistema de Gestión de Base de Datos y la lógica de negocio, mientras que el controlador es el responsable de recibir los eventos de entrada desde la vista.

La utilización de esta arquitectura garantiza que el modelo, la vista y el controlador se desarrollen de forma independiente, permitiendo que las modificaciones que se puedan realizar en alguna de estas partes no afecten a las demás. También permite potenciar el trabajo en equipo, eliminando el exceso de responsabilidades a un mismo desarrollador. Aumentando de esta forma la rapidez y calidad con la que se desarrolla el proyecto.

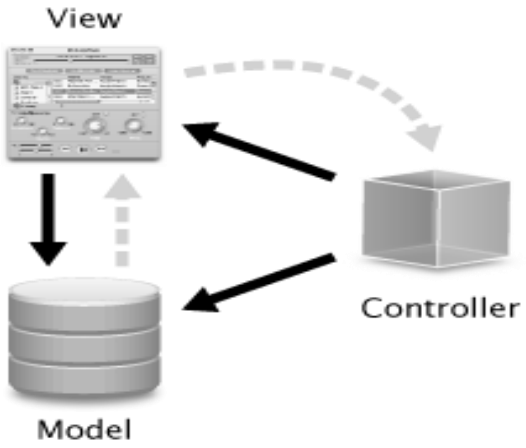


Figura 4: Modelo Vista Controlador (MVC).

3.3 Patrones de diseño.

Un patrón de diseño describe una estructura de diseño que resuelve un problema de diseño en particular. Estos patrones se aplican a un elemento específico del diseño, relaciones entre componentes o los mecanismos para efectuar la comunicación de componente a componente. (Roger, 2005).

Para el desarrollo de la solución se aplicaron los patrones de diseño, Patrones Generales de Asignación de Responsabilidad de Software (GRASP por sus siglas en inglés) y la Banda de los Cuatro (GoF por sus siglas en inglés).

Los patrones GRASP describen los principios fundamentales de la asignación de responsabilidades a objetos, expresados en formas de patrones. El nombre se eligió para indicar la importancia de captar estos principios, con el objetivo de diseñar eficazmente el software orientado a objetos. (Roger, 2005).

Experto: Permite asignar responsabilidades entre las clases, estableciendo para cada clase su responsabilidad de acuerdo a la información que posee. Dicho patrón se utilizó en el sistema para definir

las funcionalidades a desarrollar por cada clase dando una mejor comprensión del código pensado a desarrollar, específicamente en la clase *modelo.php* y *vista.php* ya que ellos son los que conocen la información que manejan.

Creador: El patrón Creador guía la asignación de responsabilidades relacionadas con la creación de objetos, tarea muy frecuente en los sistemas orientados a objetos. El propósito fundamental de este patrón es encontrar un creador que se debe conectar con el objeto producido en cualquier evento. (Larman, 1999). Permite asignar responsabilidades a las clases admitiendo la creación de instancias de las mismas. El mismo se evidencia en las clases *controladora.php,index.php y noticia_completa.php* encargadas de crear instancia sobre *vista.php* y *controladora.php* respectivamente.

Controlador: Sirve de intermediario entre una clase interfaz y la clase que contiene el algoritmo de la funcionalidad. Todas las peticiones Web son manejadas por un solo controlador frontal, que es el punto de entrada único de toda la aplicación en un entorno determinado. La constancia de este patrón en el desarrollo del sistema se pone de manifiesto en la clase *controladora.php*, esta clase se encarga de manejar todo el flujo de información del sistema.

Bajo Acoplamiento: El bajo acoplamiento se utiliza en la creación de las clases encargadas de la lógica del negocio, donde cada una se corresponde con una funcionalidad específica, lo que permite su reutización en varios negocios. En el sistema se evidencia su uso en las clases que componen el paquete de procesamiento, identificadas por los requisitos funcionales.

Los patrones GoF se clasifican en tres grupos principales (estructurales, creacionales y de comportamiento). Los patrones creacionales abstraen el proceso de instanciación, procuran independizar el sistema de cómo sus objetos son creados, compuestos y representados. Los patrones de comportamiento permiten conocer no solo aspectos deseables del sistema de índole estructural sino especialmente sus características dinámicas. Por último los patrones estructurales se centran en problemas relacionados con la forma de estructurar las clases. (Roger, 2005). Dentro de los patrones estructurales se utilizó:

Fachada: Provee de una interfaz unificada simple para acceder a una interfaz o grupo de interfaces de un subsistema. En el sistema se utiliza en la comunicación con el motor de búsqueda de Internet Google, a través de un Protocolo de Interfaz de Acceso (API por sus siglas en inglés) para acceder solamente a la búsqueda de noticias.

3.3.1 Diagrama de Clases del Diseño.

Los Diagramas de Clases del Diseño son los diagramas principales en el flujo de trabajo análisis y diseño para obtener un sistema. Durante el diseño del sistema, el diagrama de diseño se desarrolla enfocado

a los detalles de la implementación. El diagrama de clases del diseño muestra el sistema en términos de clases y métodos. Refleja el funcionamiento de la aplicación en términos lógicos. (Moya, 2011). A continuación se muestra el diagrama de clases del diseño correspondiente al funcionamiento del sistema.

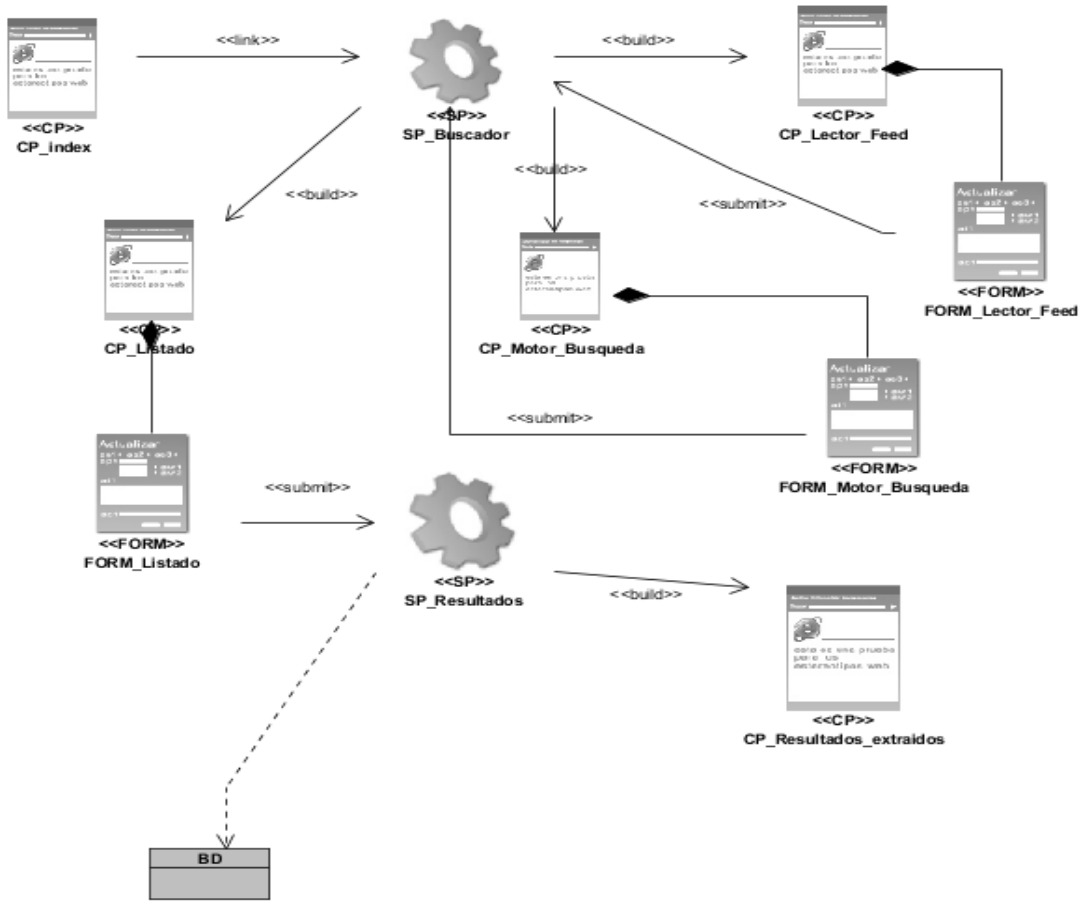


Figura 5: Diagrama de clase del diseño.

El mismo está compuesto por una página cliente inicial denominada CP_Index, la cual accede a una página controladora denominada SP_Buscador, la cual se encarga de interpretar la petición del usuario y proveerle la interfaz adecuada de acuerdo a su solicitud las cuales pueden ser, la CP_Lector_Feed, que se encarga de manejar todo lo referente a la gestión de las fuentes web y la CP_Motor_Busqueda que se encarga de hacer una búsqueda en Internet sobre un tema específico escogido por el usuario. Ambas interfaces manejan el tema de recuperación de la información noticiosa y cuentan con los formularios, FORM_Lector_Feed y FORM_Motor_Busqueda respectivamente, que se encargan de gestionar la información correspondiente al tipo de solicitud realizada por el usuario haciendo un envío de datos a la misma página controladora para visualizar una interfaz denominada CP_Listado la cual contiene las noticias recuperadas ya sea por la CP_Lector_Feed o por la CP_Motor_Busqueda. La CP_Listado tiene asociado un formulario denominado FORM_Listado el cual se encarga, de acuerdo a

la noticia seleccionada por el usuario para realizar la extracción de sus datos, de realizar el envío de estos a otra página controladora denominada SP_Resultados la cual muestra la interfaz CP_Resultados_extraidos que contiene los datos extraídos referente a la noticia recuperada, dejando listo el escenario para realizar el almacenamiento de los mismo en un Sistema Gestor de base de Datos denominado BD.

3.4 Diagrama Entidad-Relación.

El modelo de datos es una representación abstracta de los datos de una organización y las relaciones entre ellos. El propósito de un modelo de datos es, por una parte, representar los datos y, por otra, ser comprensible. Elaborar un correcto modelo de datos es imprescindible para manejar y obtener correctamente los datos. (Rosa-Rosario., 2011). A continuación se muestra el modelo de datos que representa físicamente la base de datos. Esta fue diseñada para almacenar todos los datos referentes al proceso de extracción del contenido de la noticia.

Figura 6: Diagrama Entidad-Relación.

Entre las tablas existentes en la base de datos, destinadas a manipular todos los datos referentes a la noticia se encuentra la tabla *noticia* encargada de recoger todos los elementos de la noticia que han sido recuperados y extraídos, de la misma se guarda su título, resumen, fuente, fecha de publicación, cuerpo, autor, subtitulo y el identificador de la fuente web que pertenece si la noticia fue procesada por el lector de fuentes web, en caso de haber sido procesada por el motor de búsqueda este campo es nulo; la tabla *feed_rss* representa los datos obtenidos a la hora de adicionar una fuente web, como su nombre, url y si necesita proxy o no para su conexión a Internet; otra de las tablas es referente a la *imagen*, la cual, al

igual que la tabla *video* y la tabla *audio* almacenan la url de la que procede el archivo y el identificador de la noticia a la que pertenece.

3.5 Diagrama de Componentes.

Un diagrama de componentes y la estructura del sistema en ejecución. Un diagrama de componentes muestra las dependencias lógicas entre componentes software, sean éstos componentes fuentes, binarios o ejecutables.

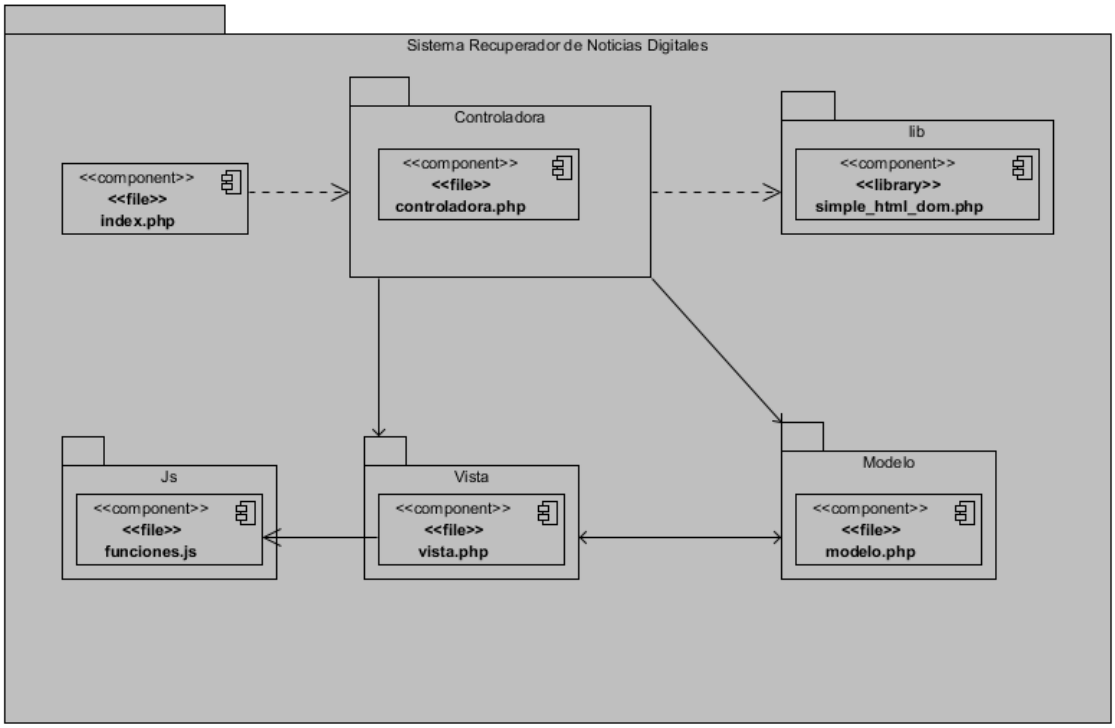


Figura 7: Diagrama de Componentes.

El diagrama anterior muestra la distribución e interacción entre los componentes que conforman el Sistema Recuperador de Noticias Digitales, teniendo en cuenta el diseño y la arquitectura establecida para el mismo.

En el paquete **Vista** se encuentra la clase **vista.php**, la cual conjuntamente con el archivo **index.php** representan la interfaz principal para realizar la interacción cliente-aplicación. **Vista.php** utiliza el archivo funcione.js dentro del paquete **Js**, el cual contiene toda la programación del lado del cliente.

Dentro del paquete **Controladora** se encuentra la clase **controladora.php**, la cual contiene todas las funcionalidades del sistema. La misma para la realización de su función depende de la librería **simple_html_dom.php** contenida dentro del paquete **lib**.

El paquete **Modelo** contiene la clase **modelo.php**, la cual se encarga de interactuar con la base de datos.

La controladora accede a la vista y al modelo, mientras estas dos se relacionan entre ellas, evidenciándose así el patrón Modelo-Vista-Controlador.

3.6 Diagrama de Despliegue.

El Diagrama de despliegue muestra las relaciones físicas de los distintos nodos que componen un sistema y el reparto de los componentes sobre dichos nodos, la configuración del sistema incluyendo software y hardware.

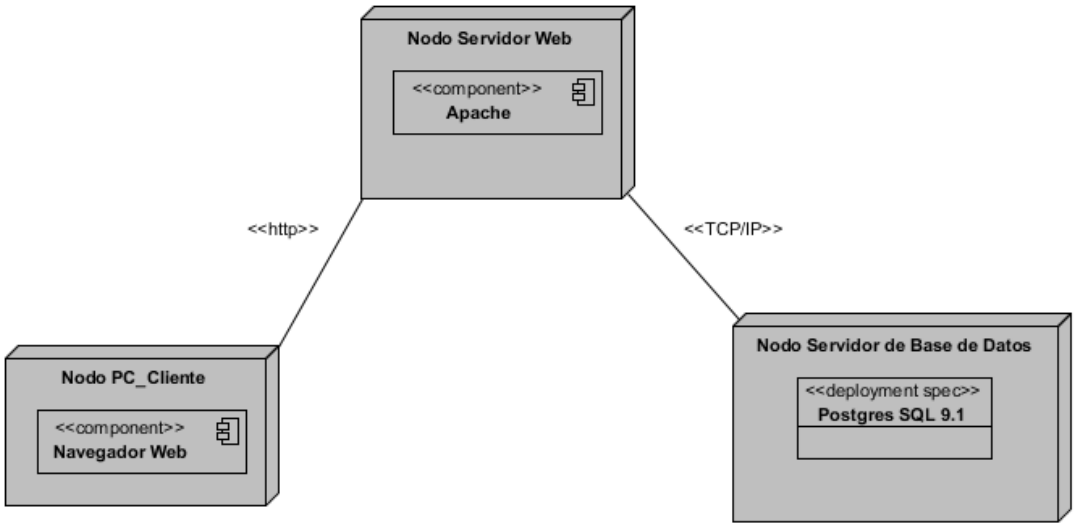


Figura 8: Diagrama de despliegue.

Base de Datos: Es el nodo encargado de contener la base de datos, además de responder a las peticiones de los clientes.

Servidor Web/PC Cliente: Nodo cliente que interactúa con la plataforma.

HTTP: Se utilizará el protocolo de comunicación HTTP ya que el sistema está implementado en tecnología Web.

3.7 Conclusiones parciales.

El uso de patrones permitió elaborar diagramas de clases de diseño fiables para la implementación de los diferentes casos de uso del Sistema Recuperador de Noticias Digitales. La elaboración de los diagramas de clases del diseño, junto a otros artefactos como el modelo de datos, el diagrama de despliegue, y el modelo de implementación, sirvió de guía para la implementación del sistema y servirá de base para que otros desarrolladores mejoren sus funcionalidades y le den mantenimiento al software.

Capítulo 4: Validación de la solución propuesta.

4.1 Introducción.

Las pruebas del software son un elemento crítico para garantizar la calidad de la aplicación desarrollada. Representan una revisión final de las especificaciones, del diseño y de la codificación de dicha aplicación. Tienen como objetivo general validar un programa, independientemente de las características y el entorno donde se desarrolla, además permite documentar los errores en el sistema y posibilita probar la correcta implementación de los requisitos que el software debe cumplir.

En la actualidad existen varios tipos de prueba que su función fundamental es garantizar la calidad del software desarrollado y detectar los errores que este puede presentar. Un ejemplo sería las pruebas de caja negra, las cuáles se encargan de estudiar el comportamiento de la aplicación, teniendo en cuenta los datos de entrada y los de salida, desconociéndose el funcionamiento interno del programa.

Estas pruebas se centran en lo que se espera de la aplicación, es decir, intentan encontrar casos en que el módulo no funcione correctamente. Es por eso que se denominan pruebas funcionales, y el probador se limita a suministrarle datos de entrada y estudiar la salida, sin preocuparse de lo que pueda estar haciendo el módulo por dentro.

4.2 Pruebas de Caja Negra.

Estas pruebas permiten obtener un conjunto de condiciones de entrada que ejerciten completamente todos los requisitos funcionales de un programa. En ellas se ignora la estructura de control, concentrándose en los requisitos funcionales del sistema y ejercitándolos. Muchos autores consideran que estas pruebas permiten encontrar (ECURED, 2012):

1. Funciones incorrectas o ausentes.
2. Errores de interfaz.
3. Errores en estructuras de datos o en accesos a las Bases de Datos externas.
4. Errores de rendimiento.
5. Errores de inicialización y terminación. En resumen, las pruebas de caja negra permiten encontrar errores como funciones incorrectas o ausentes, en la interfaz, en la estructura de datos, de rendimiento y de inicialización y terminación.

4.3 Casos de Prueba.

La realización de las pruebas de caja negra al Sistema Recuperador de Noticias Digitales, requirió el diseño de los casos de prueba en correspondencia con las descripciones de casos de uso. Los casos de prueba permiten comprobar todos los flujos de información del sistema y validar que este cumple con los requisitos del cliente. A continuación se presentan los casos de prueba utilizados para el caso de uso Buscar Noticias, un resumen del resto de los casos de prueba de los casos de usos principales se puede consultar en el **Anexo 2**.

4.3.1 Caso de Prueba para el Caso de Uso Buscar Noticias.

Tabla 4: Diseño de caso de prueba de caja negra del caso de uso Buscar Noticias.

Caso de prueba					
Nombre de la sección	**Escenarios de la sección**	**Descripción de la funcionalidad**	**Flujo central**	**Resultados esperados**	**Resultados obtenidos**
SC1: Buscar noticias.	EC1.1: Buscar noticias exitosamente.	El usuario introduce el tema a buscar en el campo de búsqueda y presiona el botón "Buscar".	Se accede a la aplicación y en el buscador se selecciona la opción "Buscar."	El sistema debe mostrar un listado de titulares de las noticias relacionadas con el criterio de búsqueda insertado por el usuario.	Satisfactorio.
	EC1.2: Dejar campo vacíos.	El usuario deja el campo de búsqueda vacío y el sistema muestra un mensaje de error.		El sistema debe indicar que el campo está vacío y muestra un mensaje de error indicándolo.	Satisfactorio.
	EC1.3: Buscar noticia fallida.	El usuario introduce un tema de búsqueda que no existe y el sistema muestra un mensaje indicando que no se encontraron resultados para esa búsqueda.		El sistema debe indicar un mensaje de error reflejando que no se encontraron resultados para la búsqueda realizada por el usuario.	Satisfactorio.

4.4 Resultado de las Pruebas.

El resultado de las pruebas realizadas al software fue satisfactorio, se comprobó que las respuestas del sistema son las esperadas y que el mismo cumple con las especificidades planteadas y a su vez coinciden con las descripciones de los casos de uso planteados con anterioridad. Las validaciones propuestas confirmaron el exitoso funcionamiento de la aplicación, se comprobaron sus validaciones con datos erróneos y datos válidos, donde por cada variable o campo a evaluar, se obtuvieron resultados satisfactorios. Con las pruebas aplicadas se comprobó que el sistema cumple con los objetivos propuestos.

4.5 Conclusiones Parciales.

Las pruebas realizadas a la aplicación lograron detectar y documentar a tiempo los errores que existían para poder corregirlos. Además validaron que la implementación satisface todos los requisitos identificados. Al finalizar con la fase de prueba se obtuvo como resultado una aplicación que funciona correctamente, cumpliéndose así la totalidad de los objetivos planteados para Sistema Recuperador de Noticias Digitales. Dicha aplicación está lista para ser desplegada.

Conclusiones Generales

Después de haber desarrollado y analizado los resultados obtenidos con la elaboración del presente trabajo y la culminación del diseño de las funcionalidades del Sistema Recuperador de Noticias Digitales, se llegaron a las siguientes conclusiones:

- Con el empleo de los métodos de investigación teóricos se logró conocer el estado del objeto de estudio de la investigación. Analizando los principales conceptos asociados al tema de investigación y las características presentadas por tres de los motores de búsquedas consideradas más importantes; así como la identificación de los procesos y funcionalidades presentas en la redifusión de contenido web, se comprendió de manera satisfactoria toda la situación problemática planteada.
- Como parte de las tareas a cumplir en el presente trabajo se desarrolló el levantamiento de requisitos y el modelado del dominio existente de una manera exitosa, el cual permitió identificar los principales conceptos y sus relaciones, involucrados en el contexto del problema, dando la posibilidad de identificar los casos de uso del sistema como elemento clave en el desarrollo de software apoyado en la metodología RUP.
- Con el planteamiento de la arquitectura se pudo estructurar los requisitos definidos previamente constituyendo un primer acercamiento al diseño y logrando su posterior realización a través del diagrama de clase del diseño obteniendo como resultado el modelado del diagrama de despliegue para tener así una visión clara de cómo se va a desplegar el sistema propuesto.

Por todo lo anterior se concluye que los objetivos propuestos para el presente trabajo han sido cumplidos satisfactoriamente.

Recomendaciones

Una vez cumplido con los objetivos del presente trabajo y en correspondencia con los resultados obtenidos, los cuales se pusieron en práctica en el propio documento, se recomienda:

1. Continuar desarrollando la investigación para perfeccionar los algoritmos referentes a la extracción de los elementos *audio* y *video*.
2. Continuar perfeccionando el algoritmo para la obtención del formato texto de la noticia.

Trabajos citados

Baeza-Yates. 2000. *Retrieve Information .* 2000.

Chang, G. 2001. *Mining the World Wide Web: an information search approach".* 2001.

Copr., IBM. 2006. *Rational Unified Process Versión 7.0.1.* 2006.

Craig, Larman. 1999. *Introducción al analisis y diseño orientado a objetos.* Mexico : s.n., 1999.

ECURED. 2012. http://www.ecured.cu/index.php/Pruebas_de_caja_negra. [En línea] 12 de 5 de 2012.

Eguiluz, Javier. 2009. *Introduccion a CSS.* 2009.

Fuentes, Yoandry. 2007. *Flujo de trabajo de requerimientos.* La Habana : s.n., 2007.

Gonzalez, Carlso. 2010. http://www.enterate.unam.mx/Articulos/rss.htm. [En línea] 2010.

Hernandez, Carlos. 2011. *Diferentes lenguajes de programacion para la web.* 2011.

Hernandez, Javier. 2011. http://www.articulandia.com/premium/article.php/17-04-2007Analisis-y-Caracteristicas-de-Google-Parte-1.htm. [En línea] 2011.

Hill, Salton Mc Graw. 1983. *Introducttion to Modern Information Retrieval.* New York : s.n., 1983.

Introducción a CSS. **Pérez, Javier Eguíluz. 2009.** 2009.

Johnson. 2006. *RSS and Atom in Action. .* 2006.

Korfhage. 1997. *Information Storage and Retrieval.* New York : s.n., 1997.

Kyrnin. 2008. *What is RSS and how do you use it.* 2008.

Larman, Craig. 1999. *UML Y PATRONES.* Mexico : s.n., 1999.

Mani, I., y otros. 1999. *"The TIPSTER SUMMAC Text Summarization Evaluation". En: Proceedings of EACL´99. 1999.* Bergen,Noruega : s.n., 1999.

Martinez, J.L. 1962. *Guiones de clase de Redacción Periodística.* Pamplona : s.n., 1962.

METABUSCADORES. 2008. http://iso-go.es/2010/09/metabuscadores-concepto-de-uso-ventajas-e-inconvenientes-2/. [En línea] 2008.

Moya, Iván Hernández. 2011. *Migración del Módulo de Catalogación de materiales audiovisuales del producto Captura y catalogación de medias.* La Habana : s.n., 2011.

Pérez, M. 1999. *Arquitectura para Ambientes CASE Integrados. .* Universidad Central de Las Villas : s.n., 1999.

PostgreSql. **2012.** 2012.

2011. *programacionya. programacionya.* 2011.

R., Grisham. 1997. *Information Extraction: Techniques and Challenges.* 1997.

Roger, Pressman. 2005. *Ingeniería de Software.Un enfoque práctico. .* Espana : s.n., 2005.

Rosa-Rosario., Dra. María G. 2011. Modelos de bases de datos. *Base de Datos Relacionales.* 2011.

Sayre, Nottingham. 2005. *The Atom Syndication Format - RFC 4287.* 2005.

Shabb, Veronica. 2004. *Verónica. Formatos periodísticos para prensa escrita.* 2004.

T.A.S, Pardo. 2003. *GistSumm: A Summarization Tool Based on a New Extractive Method.* 2003.

Veluchamy, Thiyagarajan. 2010. *IEEE 610. .* 2010.

W.B, Crofft. 2000. *Approaches to intelligent information retrieval.* 2000.

W3C. 2006. *W3C Consortium, XML Specification.* 2006.

WIKI. 2010. www.wikipedia.org.es/Bing. [En línea] 2010.

WIKIPEDIA. 2013. (http://es.wikipedia.org/wiki/Atom_%28formato_de_redifusi%C3%B3n%29). [En línea] 2013.

Yee. 2009. *Pro Web 2.0 Mashup: Remixing Data and Web Services. : Apress.* 2009.

Yunior. 2008. *Sistema de teletexto para plataforma de televisión digital satelital.* 2008.

Bibliografía

Baeza-Yates. 2000. *Retrieve Information .* 2000.

Chang, G. 2001. *Mining the World Wide Web: an information search approach".* 2001.

Copr., IBM. 2006. *Rational Unified Process Versión 7.0.1.* 2006.

Craig, Larman. 1999. *Introducción al analisis y diseño orientado a objetos.* Mexico : s.n., 1999.

developer.mozilla. developer.mozilla. [En línea] [Citado el: 25 de Noviembre de 2011.]https://developer.mozilla.org/es/Gu%C3%ADa_JavaScript_1.5/Concepto_de_JavaScript. [En línea]

ECURED. 2012. http://www.ecured.cu/index.php/Pruebas_de_caja_negra. [En línea] 12 de 5 de 2012.

Eguiluz, Javier. 2009. *Introduccion a CSS.* 2009.

Evaluación RI Y MW ICIMAF . **2013.** 2013.

Fuentes, Yoandry. 2007. *Flujo de trabajo de requerimientos.* La Habana : s.n., 2007.

Glosario breve descripción de terminología usada en esta página. Available from: http://www.codebox.es/glosario. [En línea]

Gonzalez, Carlso. 2010. http://www.enterate.unam.mx/Articulos/rss.htm. [En línea] 2010.

Hernandez, Carlos. 2011. *Diferentes lenguajes de programacion para la web.* 2011.

Hernandez, Javier. 2011. http://www.articulandia.com/premium/article.php/17-04-2007Analisis-y-Caracteristicas-de-Google-Parte-1.htm. [En línea] 2011.

Hill, Salton Mc Graw. 1983. *Introducttion to Modern Information Retrieval.* New York : s.n., 1983.

http://bitelia.com/2011/06/google-bing-yahoo-unidos-optimizar-indexacion-webs. [En línea]

http://blogs.msdn.com/b/expressate/archive/2009/06/03/usando-los-servicios-y-api-de-bing.aspx. [En línea]

http://datamarket.azure.com/dataset/bing/search. [En línea]

http://es.scribd.com/doc/13303308/Motores-de-Busqueda-Arquitectura-de-Google. [En línea]

http://es.wikipedia.org/wiki/Google_search_ajax_api. [En línea]

http://es.wikipedia.org/wiki/Google_SOAP_Search_API. [En línea]

http://es.wikipedia.org/wiki/RSS. [En línea]

http://wiki-tic-cervantes.wikispaces.com/METABUSCADORES. [En línea]

http://www.ayudaenlaweb.com/buscadores/que-es-yahoo/. [En línea]

http://www.ayudaenlaweb.com/buscadores/que-es-yahoo/. [En línea]

http://www.cad.com.mx/historia_de_yahoo.htm. [En línea]

http://www.developer.com/lang/article.php/3691506/Introducing-the-Google-AJAX-APIs.htm. [En línea]

http://www.hostito.com/es/faq/index.php?section=glossary&title_section=Glosario%20de%20T%C3%A9rminos#30. [En línea]

http://wwwcorreosnavegadoresbuscadores.blogspot.com/2010/06/ventajas-y-desventajas-de-yahoo.html. [En línea]

Introducción a CSS. **Pérez, Javier Eguíluz. 2009.** 2009.

Johnson. 2006. *RSS and Atom in Action. .* 2006.

Korfhage. 1997. *Information Storage and Retrieval.* New York : s.n., 1997.

Kyrnin. 2008. *What is RSS and how do you use it.* 2008.

Larman, Craig. 1999. *UML Y PATRONES.* Mexico : s.n., 1999.

Mani, I., y otros. 1999. *"The TIPSTER SUMMAC Text Summarization Evaluation". En: Proceedings of EACL´99. 1999.* Bergen,Noruega : s.n., 1999.

Martinez, J.L. 1962. *Guiones de clase de Redacción Periodística.* Pamplona : s.n., 1962.

METABUSCADORES. 2008. http://iso-go.es/2010/09/metabuscadores-concepto-de-uso-ventajas-e-inconvenientes-2/. [En línea] 2008.

Moya, Iván Hernández. 2011. *Migración del Módulo de Catalogación de materiales audiovisuales del producto Captura y catalogación de medias.* La Habana : s.n., 2011.

Pérez, M. 1999. *Arquitectura para Ambientes CASE Integrados. .* Universidad Central de Las Villas : s.n., 1999.

PostgreSql. **2012.** 2012.

2011. *programacionya. programacionya.* 2011.

R., Grisham. 1997. *Information Extraction: Techniques and Challenges.* 1997.

Roger, Pressman. 2005. *Ingeniería de Software.Un enfoque práctico. .* Espana : s.n., 2005.

Rosa-Rosario., Dra. María G. 2011. Modelos de bases de datos. *Base de Datos Relacionales.* 2011.

Sayre, Nottingham. 2005. *The Atom Syndication Format - RFC 4287.* 2005.

Shabb, Veronica. 2004. *Verónica. Formatos periodísticos para prensa escrita.* 2004.

T.A.S, Pardo. 2003. *GistSumm: A Summarization Tool Based on a New Extractive Method.* 2003.

Veluchamy, Thiyagarajan. 2010. *IEEE 610. .* 2010.

Visual paradigm. Visual-paradigm. Visual-paradigm. Available from: http://www.visual-paradigm.com/. [En línea]

W.B, Crofft. 2000. *Approaches to intelligent information retrieval.* 2000.

W3C. 2006. *W3C Consortium, XML Specification.* 2006.

WIKI. 2010. www.wikipedia.org.es/Bing. [En línea] 2010.

WIKIPEDIA. 2013. (http://es.wikipedia.org/wiki/Atom_%28formato_de_redifusi%C3%B3n%29). [En línea] 2013.

Yee. 2009. *Pro Web 2.0 Mashup: Remixing Data and Web Services. : Apress.* 2009.

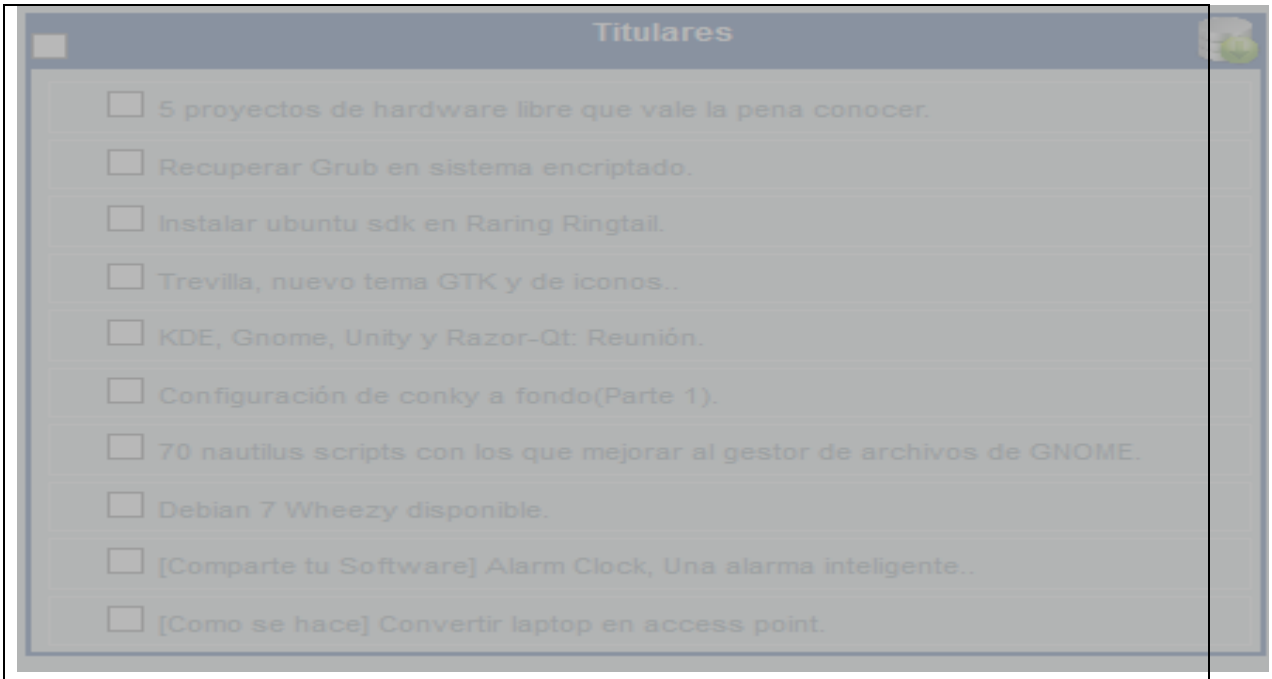
Yunior. 2008. *Sistema de teletexto para plataforma de televisión digital satelital.* 2008.

Anexo 1

Tabla 5: Especificación del CU Manipular Noticia.

Caso de Uso: CU-1	Manipular Noticia.
Actores:	Redactor.
Propósito:	Permitir que se pueda mostrar los titulares de cada una de las noticias pertenecientes a una fuente web seleccionada.
Resumen:	El CUS se inicia cuando el Redactor selecciona la fuente web deseada a consultar. Después de consultada el sistema muestra el resultado. Una vez realizada la operación termina el CUS.
Referencias:	RF4.
Precondiciones:	El Redactor debe seleccionar una fuente web.
Pos condiciones:	Se muestra el título de todas las noticias.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El Redactor selecciona la opción Adicionar RSS.	2. El sistema muestra nombre y URL y si requiere proxy para la fuente web a adicionar.
3. El Redactor llena los datos y selecciona la opción Agregar.	4. El sistema muestra la(s) fuente(s) web agregada(s).
5. El Redactor selecciona la fuente web a consultar.	6. El sistema muestra los titulares más actualizados pertenecientes a la fuente web seleccionada. Una vez realizada la operación termina el CUS.



Flujo Alterno.	
1. El Redactor selecciona una fuente web no disponible.	2. El sistema muestra un mensaje de confirmación indicando la fuente web no está disponible en ese momento y no puede listar las noticias. Termina el escenario

Tabla 6: Especificación del CU Buscar Noticia.

Caso de Uso: CU-2	Buscar Noticia.
Actores:	Redactor.
Propósito:	Permitir buscar información de carácter noticioso solamente a través del motor de búsqueda de Google.
Resumen:	El CUS se inicia cuando el Redactor inserta el tema de interés a consultar y selecciona la opción Buscar. Después de realizada la búsqueda el sistema muestra el resultado. Una vez realizada la operación termina el CUS.
Referencias:	RF5, RF6.
Precondiciones:	El Redactor debe ingresar el tema para realizar la búsqueda.

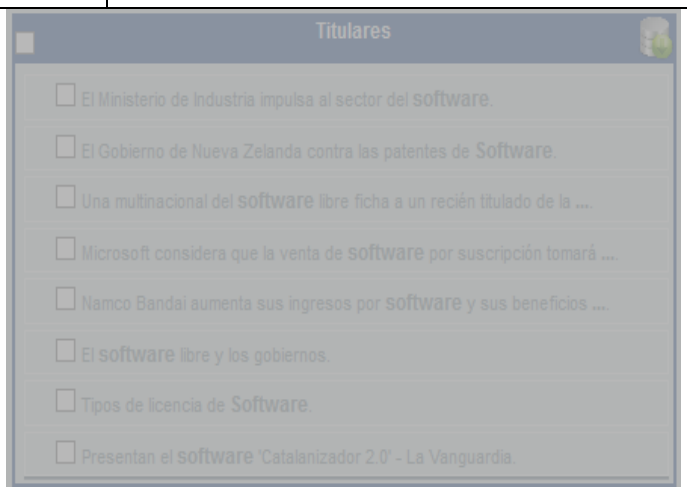
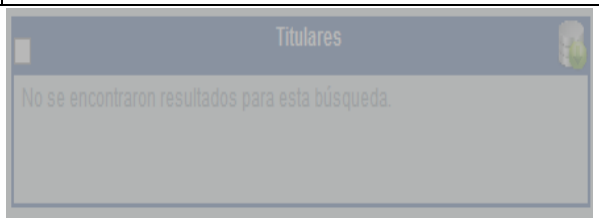
Pos condiciones:	Se muestra el resultado de la búsqueda mediante un listado.
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El Redactor ingresa el tema noticioso que desea consultar y selecciona la opción Buscar.	2. El sistema muestra un listado con los titulares del tema consultado resaltando en *negrita* el mismo.
	
Flujo Alterno.	
1. El Redactor ingresa un tema noticioso que no existe.	2. El sistema muestra un mensaje indicando que no es encontraron resultados para esa búsqueda.
	

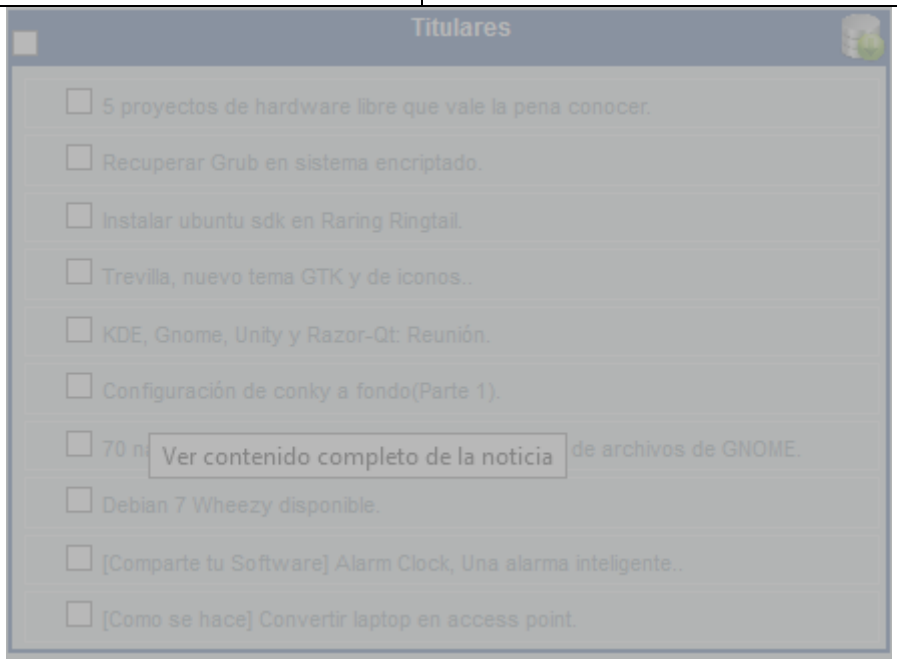
Tabla 7: Especificación del CU Extraer Contenido.

Caso de Uso: CU-3	Extraer Contenido.
Actores:	Redactor.
Propósito:	Permitir extraer el contenido completo de la noticia.

Resumen:	El CUS se inicia cuando el Redactor selecciona la noticia de la cual quiere ver su contenido completo. Después de realizada la operación el sistema muestra el resultado. Una vez realizada la operación termina el CUS.
Referencias:	RF7, RF8.
Precondiciones:	El Redactor debe seleccionar una noticia.
Pos condiciones:	Se muestra la fuente a la que pertenece la noticia, su fecha de publicación, titulo, sub-título, autor, resumen, cuerpo de la noticia y sus elementos asociados.

Flujo Normal de Eventos

Acción del Actor	Respuesta del Sistema
1. El Redactor selecciona fuente web para ver sus titulares.	2. El sistema muestra todas las noticias pertenecientes a la fuente web seleccionada.
3. El Redactor selecciona una noticia para ver su contenido.	4. El sistema muestra los datos extraídos de la noticia seleccionada. Una vez realidad la acción termina el CUS.



Noticias

Fuente: humanos.uci.cu.

Fecha: Mon, 06 May 2013 12:05:08 +0000.

Titulo: Configuración de conky a fondo(Parte 1).

Sub-Titulo: Desconocido.

Autor: Desconocido.

Resumen: Desconocido.

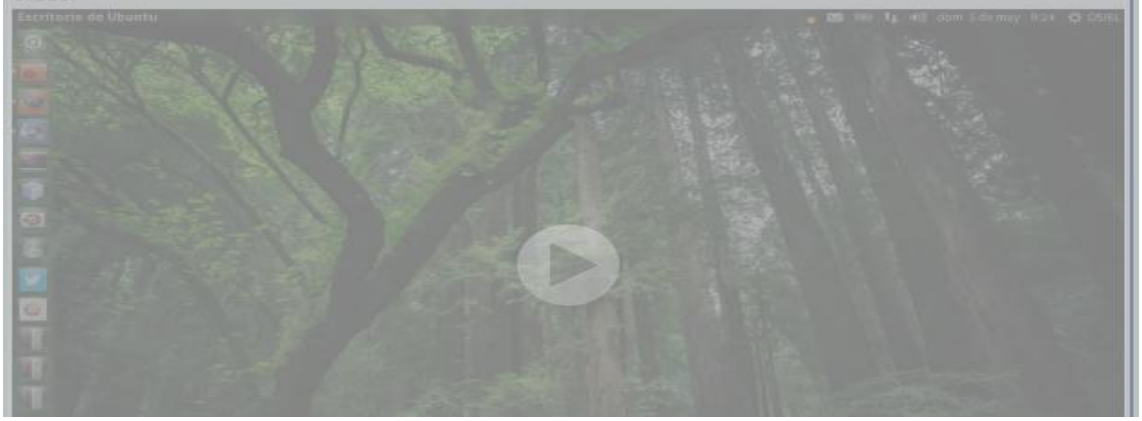
Cuerpo:

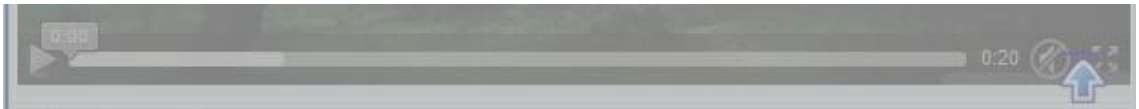
Mucho se ha hablado del conky, esa maravillosa herramienta para mostrar información en nuestro escritorio a gusto propio. Acá en el blog hemos dado seguimiento a esta herramienta enseñando como iniciarnos en ella, pero una de las cosas que me gustan es la cantidad de información que puedes mostrar y lo fácil de configurar que es. En el siguiente video, podemos observar algo pequeño de lo que lograremos al terminar este grupo de tutoriales sobre el tema.
 ,Les propongo hacer un recorrido básico de las principales cosas que se debemos conocer para instalar y poner a correr conky y después centrarnos en la configuraciones que podemos realizar.,Primeramente tendremos que instalar conky, utilizaremos para ello la herramienta apt y con el comando siguiente podremos instalarlo desde el repositorio:,También podemos instalarlo utilizando el Centro de software o el Synaptic en Ubuntu, o haciendo click aquí en Debian, Ubuntu y derivados que utilicen APT.,Una vez

Elementos asociados:

Imagen(s):

Video:





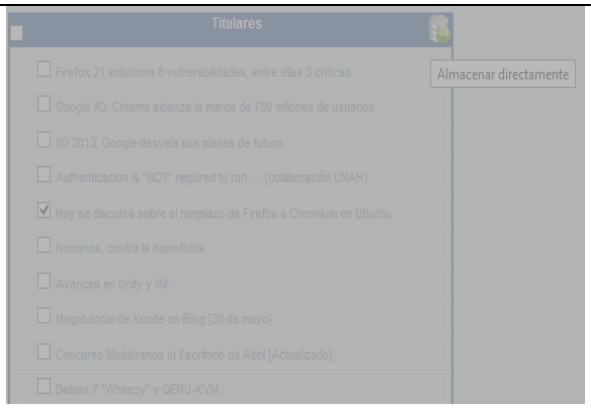
Audio: Desconocido.

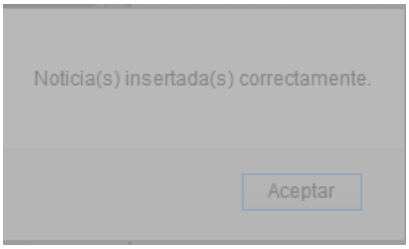
	1. Cuando una noticia no está disponible para su procesamiento el sistema muestra un mensaje reflejando que no es una URL valida. Una vez realizada la operación termina el CUS.

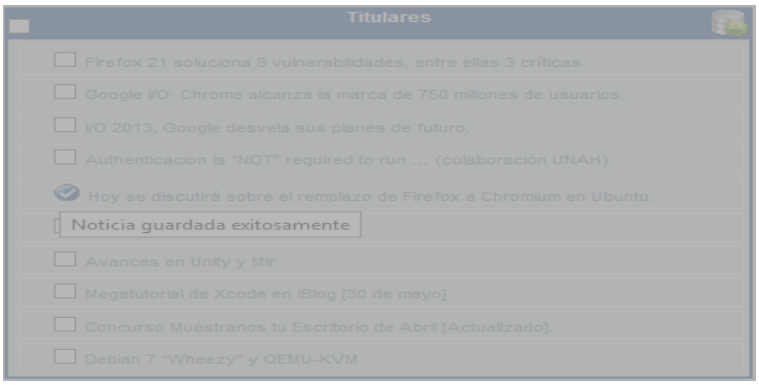
Tabla 8: Especificación del CU Almacenar Contenido Directo.

Caso de Uso: CU-4	Almacenar Contenido Directo.
Actores:	Redactor.
Propósito:	Permitir almacenar la información extraída en la base de datos sin tener que consultar su contenido.
Resumen:	El CUS se inicia cuando el Redactor selecciona la opción Almacenar directamente. Una vez realizada la operación termina el CUS.
Referencias:	RF10, RF11.
Precondiciones:	El Redactor debe seleccionar la(s) noticia(s) para guardar.
Pos condiciones:	Se muestra un mensaje indicando que la noticia ha sido guardada exitosamente.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El Redactor selecciona la fuente web de la cual quiere almacenar su(s) noticia(s).	2. El sistema muestra un listado con los titulares de la noticia.
3. El Redactor selecciona la opción de marcar todo o marcar una sola noticia y selecciona la opción “almacenar directamente.”	4. El sistema almacena correctamente la noticia sin mostrar su contenido y muestra un mensaje de confirmación indicando el éxito. Una vez realizada la acción termina el CUS.







Flujo Alterno.	
1. El Redactor no selecciona la(s) noticia(s) para guardar.	2. El sistema muestra un mensaje indicando que tiene que seleccionar la(s) noticia(s) a guardar.

Anexo 2

Tabla 9: Diseño de caso de prueba de caja negra del caso de uso Manipular Noticias.

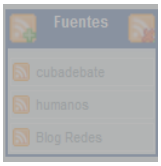
Caso de prueba					
Nombre de la sección	**Escenarios de la sección**	**Descripción de la funcionalidad**	**Flujo central**	**Resultados esperados**	**Resultados obtenidos**
SC1: Manipular noticias.	EC1.1: Listar las noticias pertenecientes a una fuente web.	El usuario agrega una nueva fuente web o selecciona alguna de las existentes en el sistema para ver sus noticias y se muestra un listado de las mismas.	Se accede a la aplicación y en el menú “Fuentes” se selecciona la opción “Adicionar RSS” o se selecciona una fuente web. 	El sistema debe mostrar un listado de titulares de las noticias pertenecientes a la fuente web seleccionada por el usuario.	Satisfactorio.
	EC1.2: No listar las noticias pertenecientes a una fuente web	El usuario selecciona una fuente web no disponible en ese momento.		El sistema debe indicar que dicha fuente web no se puede consultar mostrando un mensaje de error.	Satisfactorio.

Tabla 10: Diseño de caso de prueba de caja negra del caso de uso Extraer contenido.

Caso de prueba					
Nombre de la sección	**Escenarios de la sección**	**Descripción de la funcionalidad**	**Flujo central**	**Resultados esperados**	**Resultados obtenidos**
SC1: Extraer contenido.	EC1.1: Extraer el contenido de una noticia exitosamente.	El usuario selecciona la noticia listada de la cual desea ver su contenido ya sea por la fuente web o por el buscador.	Se accede a la aplicación, se selecciona una fuente web para ver sus noticias y se selecciona la noticia para ver su contenido. 	El sistema debe mostrar los datos extraídos referentes a la noticia seleccionada por el usuario.	Satisfactorio.
	EC1.2: Extraer el contenido de una noticia fallida.	El usuario selecciona una noticia que su contenido no puede ser procesado.		El sistema debe indicar que dicha URL no se puede procesar y muestra un mensaje de error indicándolo.	Satisfactorio.

Tabla 11: Diseño de caso de prueba de caja negra del caso de uso Almacenar contenido directo.

Caso de prueba					
Nombre de la sección	**Escenarios de la sección**	**Descripción de la funcionalidad**	**Flujo central**	**Resultados esperados**	**Resultados obtenidos**
SC1: Almacenar contenido directo.	EC1.1: Dejar campo desmarcado.	El usuario no selecciona la(s) noticia(s) a guardar.	Se accede a la aplicación, se selecciona la fuente web de la cual se quieren ver sus noticias y se selecciona la opción “Almacenar directamente”. 	El sistema debe mostrar un mensaje de error indicando que se debe seleccionar al menos una noticia.	Satisfactorio.
	EC1.2: Almacenar el contenido directo de una noticia exitosamente.	El usuario selecciona almacenar los datos de la(s) noticia(s) sin querer ver su contenido.		El sistema debe almacenar la(s) noticia(s) seleccionada por el usuario satisfactoriamente y mostrar un mensaje de confirmación.	Satisfactorio.
	EC1.3: Almacenar el contenido directo de una noticia sin éxito.	El usuario selecciona una noticia que no se puede procesar en ese momento y selección la opción “Almacenar directamente”.		El sistema debe mostrar un mensaje de error reflejando que esa noticia no se puede almacenar y que debe recargar la página.	Satisfactorio.