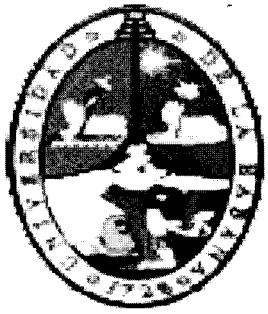
UNIVERSIDAD DE LA HABANA
DEPARTAMENTO DE COMPUTACIÓN



Trabajo de Diploma

Título

Proveedor Automático de Noticias

Autor

Yadira Ruíz Constanten

Tutor

Ing. William Azcuy Morales

CIUDAD DE LA HABANA
JUNIO 2004

Pensamiento

"No hacen falta alas para hacer un sueño,

basta con las manos, basta con el pecho,

basta con las piernas, y con el empeño".

Silvio Rodríguez

Dedicatoria

A mi Ceiba y mi Paloma.

A mis muertos...

Agradecimientos

Al ser más sabio y fuerte en su inmensa ternura cual ceiba de algodón de azúcar. A mi mamita, hombro seguro en cada golpe.

A Larisa, mi hermana, la paloma torcaza inquieta y libre. A su cariño.

A los que no están. A los que hoy no pueden darme un beso. A los que extrañaré siempre.

A mi todo, por todo. Por quererme y aceptarme como soy.

A mi pequeña pero unida familia.

A mis amigos de noches en vela, madrugadas y amaneceres. Todos lo que en septiembre de 1999, se unieron a mi camino. Más que amigos, a mis hermanos: al cuñi más lindo del mundo, a mi negro de altas y bajas, a mi guru, a mi socio Ediber, al Rolo, a Cía, por levantarme en mis caídas y hacerse cómplice de mis alegrías, a Billy, Yaime, Nara y Yaima, al team de santiagueras y su apoyo incondicional. A los amigos nuevos, los villaclareños, por hacer de este, mi último año de estudiante, uno de los más especiales: a Yailen y Suny por la paciencia y el empuje, por convertirse en otras dos hermanas.

A mis amigas de siempre Irilis, Nora, Yinia y Yanelvis. A mi hermano Reynier.

A mis profes y su empeño.

A mi tutor William, por la confianza.

A mi Comandante y sus ideas, a las que estoy indiscutiblemente atada.

A todos y cada uno de los que he recibido afecto y apoyo.

Al alquimista, por demostrarme que cuando deseas algo de verdad, todo el Universo conspira para que puedas realizarlo.

Introducción

Desde hace muchos años, ya en la era moderna, cuando alguien necesitaba información de carácter científico, comercial o de entretenimiento solía encaminarse hacia una biblioteca pública, especializada o académica, en la que un bibliotecario o referencista lo orientaba; él podía también consultar los tradicionales catálogos de autor, título, materia u otro que describiera los documentos existentes; en el peor de los casos, el problema se resolvía cuando se remitía el usuario a otra biblioteca. Pero inevitablemente se produjo un crecimiento exponencial de la literatura, sobre todo científica, que aun cuando coloca, a disposición de la comunidad académica, una gran variedad de recursos, requiere de una inversión importante de tiempo y esfuerzo para su consulta, evaluación y asimilación.

El desarrollo científico y tecnológico, con su crecimiento agigantado, ha generado, entre otros fenómenos, el incremento y perfeccionamiento acelerado de las nuevas tecnologías de información y comunicación, justamente en función de un mejor registro, procesamiento, búsqueda y diseminación de la información; sin embargo, el problema para acceder sólo a la información relevante persiste.

Sin necesidad de analizar la evolución de las tecnologías de información, está claro que su resultado más importante es Internet. Si se retoma la idea inicial, puede pensarse que ahora, cuando alguien necesita realizar una búsqueda, incluso en el tema más sencillo, piensa en Internet y no en una biblioteca tradicional, y es que Internet, es como una gran biblioteca, con múltiples departamentos especializados en diferentes materias.

Afortunadamente, a la par del crecimiento de Internet se han desarrollado y perfeccionado los motores de búsqueda, dirigidos a facilitar la navegación y el hallazgo de la información necesaria.

Los motores de búsqueda representan el medio para referir la información porque son la función más utilizada en el web y por lo tanto un evidente medio de visibilidad, además la mayoría de los accesos a un sitio web se realizan a través de ellos siendo más económicos y duraderos; pero a la luz de sus dimensiones resulta cada vez más difícil encontrar la información que uno va

buscando y por lo tanto más difícil para ellos catalogarla. Por esta razón, los motores de búsqueda son, sin duda alguna, el punto de partida de un navegante que esté buscando algo, para evitar ofrecerle resultados inútiles. Puede parecer una paradoja, pero tener muchísima información equivale a no poseer ninguna.

Evidentemente, si se comparan los motores de búsqueda de hace unos años atrás con los actuales será fácil percatarse de que la cantidad de información procesada en sus bases de datos es mucho mayor, debido precisamente a que la información en la red se multiplica a diario. Por otra parte, se estima que, mientras en 1995, apenas existía una docena de motores de búsqueda, hoy se calculan en alrededor de 2000, cada uno con características diferentes, facilidades particulares, formas de funcionamiento e interfaz propia. Si bien es cierto que en el inicio los motores de búsqueda, la preocupación de los navegantes era encontrar alguno cuyo host estuviera disponible en el momento en que fuera a hacerse uso de él o simplemente saber cuál realizaría la búsqueda de manera más fácil, en la actualidad el primer problema está en identificar, seleccionar y decidirse por uno de ellos. [1]

Indiscutiblemente, aunque ellos constituyen un importante paso de avance, no son la solución al problema. El propio incremento de los motores de búsqueda, disponibles en la red, ha impuesto la necesidad, para la mayoría de los navegantes, de "hacer búsqueda de buscadores" con el fin de determinar cuál es el mejor para un determinado tema, incluso antes de formular la búsqueda que necesita para resolver su problema de investigación.

La idea que perseguimos como objetivo general es brindar a los usuarios de la Intranet Universitaria una aplicación que de forma centralizada y única obtenga noticias actualizadas que luego serán puestas en sus manos, para contribuir de esa forma, con la formación de todos y cada uno de ellos.

Siendo más específicos, podemos decir que nuestro propósito es crear un servicio de Windows que permita alimentar una base de datos a partir de contenidos ya existentes en la red; que sea capaz de manera automática, de realizar la búsqueda, descarga, procesamiento e inserción en una base de datos con las noticias acaecidas diariamente que hayan sido publicadas en los sitios CNN en español y Granma nacional. Para ello el servicio debe apoyarse en un fichero de configuración XML que hemos de crear, para acceder a la

información publicada pues en él estarán especificadas las ubicaciones exactas de los artículos en Internet. Además se mantendrá actualizada la base de datos en las que se almacenen las noticias, dando la posibilidad al usuario de que pueda acceder a noticias ocurridas con anterioridad, promoviendo el uso eficiente de la intranet como medio de información noticiosa en la Universidad.

Capítulo 1.

Fundamentación Teórica.

1.1 Introducción

Un robot o motor de búsqueda es, en sentido general un programa que atraviesa una estructura de hipertexto recuperando todos los enlaces que están referenciados allí, que luego proveerá una base de datos que a su vez brindará la información solicitada por el buscador, según el nivel de relevancia que tenga el sitio a visitar. Ellos esencialmente, siguen los enlaces desde una página hacia otra. [2]

Nuestro sistema funciona cual un robot para abastecer de noticias actualizadas la Intranet de la Universidad de las Ciencias Informáticas, y tiene sus particularidades si lo comparamos con los proveedores de noticias en el mundo, pues en este caso lo que tomaremos de las páginas no serán todos los vínculos que en ella aparezcan, sino una pequeña parte de ellos que nos llevarán a extraer la información que necesitamos para alimentar nuestra base de datos.

Durante el desarrollo de la aplicación, haremos referencia a una serie de conceptos cuya terminología formará parte inseparable de usuarios y desarrolladores. A detallarla, dedicaremos este capítulo del informe. Además destacaremos la importancia concedida a las herramientas a utilizar en la creación del sistema.

1.2 Proveedores de noticias automáticos

Se han desarrollado en el mundo, muchos proveedores de noticias, cada uno con filosofías y patrones de búsqueda diferentes, veamos una pequeña muestra de ellos, tanto hispanos como internacionales:

Hispanos:

- GLOBAL NEW(http://cgi.grippo.com.ar/mp/go.mpc?http://www.imcg.com.ar)

El primer proveedor Digital y buscador inteligente de noticias de múltiples fuentes, formatos e idiomas, seleccionadas diariamente de más de 300 diarios,

revistas y medios especializados de 33 países, con lectura en 6 idiomas y entrega en español, inglés y portugués. Pionero en ofrecer un servicio altamente confiable con información procesada, pautada y probada. Diferenciando las notas por categorías, secciones y temas, permitiendo contar con un ágil e inmediato sistema de acceso a las fuentes. Incluye un seguimiento de noticias de sectores de la industria, política y economía, y temas concretos; consultas específicas por tema, sector o país; análisis de la evolución y los resultados de la búsqueda sobre un tema determinado y una ampliación del listado de medios de consulta a pedido del cliente, de acuerdo con sus necesidades coyunturales. [3]

- ICONOCE (http://www.iconoce.com)

Eficaz buscador de noticias hispano. Con un diseño y navegación muy cuidados, rastrea diariamente noticias procedentes de más de 500 publicaciones, incluyendo no sólo los titulares, sino el texto íntegro. [3]

- PRENSA DIGITAL (http://www.prensadigital.com)

Selecciona a diario noticias de interés general y artículos publicados en periódicos digitales españoles. Dispone de varias secciones, donde se pueden encontrar los titulares de las noticias que, mediante un enlace, se conectan con el texto original. [3]

- EXCLUSIVAS.NET (http://www.exclusivas.net)

Selección de noticias y columnas de opinión publicadas por medios digitales en español, con actualización diaria. Incluye un listado de enlaces a ediciones en Internet de publicaciones hispanas e internacionales. [3]

Internacionales:

- MOREOVER (http://www.moreover.com)

Da acceso a los titulares de más de 1.500 fuentes agrupadas en 150 categorías temáticas. Se puede acceder libremente a los textos completos. [3]

- DAYPOP (http://www.daypop.com)

Sencillo pero eficaz buscador de noticias, que permite rastrear informaciones entre más de 5.700 publicaciones digitales y weblogs (tablones personales de noticias). Ofrece la posibilidad de filtrar los resultados por idioma y por país. [3]

Hemos hecho referencia a algunos motores de búsqueda, pero debemos señalar que, de la misma forma que Google se ha convertido en un buscador por excelencia, su servicio de noticias también está dentro de los más cotizados. Veamos una breve descripción de su funcionamiento:

- Google News

Google News es una especie de periódico digital muy peculiar: tiene una página (la portada), solo se publica en Internet, es totalmente gratis e incluso carece de publicidad. Se trata de un servidor automático que recopila noticias a través de Internet, las clasifica automáticamente, las ordena por relevancia y ofrece una edición-resumen digital mundial, actualizada cada 30 minutos por lo que, cada vez que se conecte a esta página, encontrará artículos nuevos. Respeta la autoría, mantiene los enlaces a las páginas originales, dice hace cuánto rato fue publicada cada noticia original, e incluso cuántas otras noticias tiene relacionadas con esa. Todo eso funciona sin intervención humana, y el resultado es realmente fantástico.

La información que ofrece es recopilada de 700 fuentes de información de todo el mundo e incluye artículos publicados durante los últimos 30 días. Google ha creado un proceso de agrupación automatizado para Google News que reúne titulares y fotografías relacionados entre sí a partir de cientos de fuentes de información de todo el mundo, lo que le permitirá comprobar la forma en que distintos medios periodísticos explican la misma historia.

Google News es un sistema muy innovador, puesto que ofrece un servicio de información compilada únicamente mediante algoritmos informáticos, sin intervención humana alguna. Los titulares se seleccionan en función de la forma y el lugar donde aparecen los artículos en otras páginas de Internet.

Google no dispone de editores humanos que seleccionen o agrupen los titulares, por lo que ninguna persona decide qué artículos se publican en primer lugar. Por consiguiente, algunas veces, esto puede provocar que un artículo esté fuera de contexto. [4]

Nuestro sistema no es tan abarcador, apenas visita dos sitios noticiosos, Granma nacional y CNN en español, pero la principal diferencia que tiene con este gigantesco motor de búsqueda es que como nuestro sondeo es

específico, el servicio ya tiene determinado que artículos buscar en cada momento, y no correrá el riesgo de que dicho artículo quede mal clasificado y que luego al publicarlo, aparezca fuera de contexto.

1.3 Web Crawler

Los motores de búsqueda son la herramienta que permite al usuario encontrar, de una manera sencilla, cualquier tipo de información publicada en Internet. La información es clasificada de acuerdo a su relevancia o importancia; además son sistemas que, de forma automática, indexan una porción de los documentos residentes en la globalidad de la web y permiten localizar información a través de la formulación de una pregunta. Estos motores de búsqueda recopilan la información, gracias a uno o varios agentes de búsqueda (robots, spiders o crawlers) que recorren la web, a partir de una relación de direcciones de partida, recopilando nuevas páginas para el motor y generando una serie de etiquetas que permiten su indexación en la base de datos.

Los robots se denominan de diferente forma: *gusanos (worms), orugas (web crawlers), hormigas (web ants), arañas (spiders), bots, infobots,...* en función del modo en el que hagan su búsqueda, pues por ejemplo los *spiders*, son idénticos a un robot pues como ellos, se desplazan por la red, cual araña en su telaraña, pero tienen con respecto a ellos una mayor connotación en los medios de comunicación, los *worms* son programas que se replican y los robots no; los *web crawlers* son los robots que recopilan páginas web o información extraída de ellas para el índice de los motores de búsqueda; los *web ants* son robots cooperativos distribuidos, es decir robots repartidos por uno o varios servidores que unen sus esfuerzos para un fin común; los *bots* funcionan igual que un robot pero tiene un uso mas cercano al usuario no experto y los *infobots* son iguales que el anterior pero más especializado en la recolección de información. [5]

Nosotros centraremos la atención en los agentes de búsqueda en la web o web crawlers como comúnmente se les conoce, pues nuestra aplicación guarda cierta similitud con este tipo de motor de búsqueda, ya que se moverá por la red y de ella irá coleccionando algunos de los enlaces de las páginas que visita y luego sigue estos enlaces hacia otras páginas.

En 1993 se crea el primer web crawler y se le llamó "world wide worm" (gusano mundial); era un programa que se arrastraba entre un sitio y otro e indexaba todas las páginas guardando el contenido en una base de datos. Al nacer la WWW (World Wide Web) en 1994 las opciones para buscar información en la red eran bastante limitadas, existía Yahoo!, y... Yahoo!, pero con la llegada de nuevas tecnologías y mejores conexiones, aparecieron nuevos sistemas más potentes que recopilan toda la información de Internet. El crecimiento tan grande en la información publicada en Internet hace casi imposible que un sólo motor de búsqueda la mantenga indexada.

Cuando un web crawler visita una página, estudia el texto visible en los contenidos de varias etiquetas del código fuente de la página (etiqueta title, TD, etc.) y los hyperenlaces en su página. Nuestro sistema, ya dispondrá previamente de la ubicación exacta de la información que necesita extraer de las páginas que visite, a través de la distribución explicita que le brinda el fichero de configuración. Dependiendo de cómo se preparó el robot en el motor de búsqueda, la información es indexada y luego entregada a la base de datos.

Las bases de datos de los motores se actualizan varias veces. Una vez que exista una información en la base de datos del motor de búsqueda, este se mantendrá visitando periódicamente el sitio para recoger cualquier cambio que exista en él y asegurarse de que tiene la última información. El número de veces que realice las inspecciones, depende de como se hayan configurado sus visitas, las cuales pueden variar para cada motor de búsqueda. [6]

Existen algunas razones por las cuales las personas creen que los robots son malos para la Web, y ello a llevado a especulaciones en torno a la inconveniencia de los mismos, entre ellas tenemos que algunas implementaciones de robots pueden sobrecargar redes y servidores, además los robots son operados por humanos, quienes incurren frecuentemente en faltas durante la configuración, o simplemente no consideran las implicaciones de sus acciones. Pero al mismo tiempo la mayoría de los robots que están bien diseñados, que son operados profesionalmente, no causan problemas, y proporcionan un servicio valioso dada la ausencia de mejores soluciones que estén ampliamente difundidas y popularizadas. De manera que los robots no son inherentemente malos, ni tampoco inherentemente brillantes, y necesitan de una atención cuidadosa. [7]

Nuestro sistema tiene como rasgo común a los web crawlers tradicionales el hecho de que operan de forma automática al interactuar con las páginas, es decir no requieren interacción humana. Pero se diferencia en gran medida de ellos porque no clasifica las páginas que va visitando, en el fichero ya está explícita la clasificación de la información a extraer. He ahí donde radica la importancia de nuestro fichero de configuración.

1.4 XML

Es un lenguaje extensible de marcas que fue desarrollado por el Grupo de Trabajo XML (originalmente conocido como "SGML Editorial Review Board") formado bajo los auspicios del Consorcio World Wide Web (W3C), en 1996. Conocido como el lenguaje de marcas, ha surgido como uno de los formatos de información más aceptado hoy en día, inclusive en ocasiones es designado: "El ASCII de Internet", es un estándar para describir datos y crear etiquetas. Entre sus características esenciales tenemos la independencia de datos o la separación de los contenidos de su presentación, que permite estructurar información de forma sofisticada sobre formato de texto convencional e intercambiar fácilmente de esta forma entre sistemas. El XML, al ser un lenguaje extensible de etiquetas (extensible por que no es un formato prefijado como HTML), brinda una clase de objetos de datos llamados documentos XML y describe parcialmente el comportamiento de los programas que los procesan.

XML es, ante todo, un metalenguaje que permite diseñar un lenguaje propio de etiquetas para múltiples clases de documentos. Al ser un estándar fue posible desarrollar herramientas universales para operar con información XML, transformarla y validarla a nivel sintáctico y semántico. Su campo de aplicación es especialmente amplio en Internet; por eso se ha impuesto con mucha fuerza en los últimos años.

A pesar de su sencillez aparente, XML está transformando completamente la creación y el uso de software. El Web revolucionó la comunicación entre usuarios y aplicaciones. XML está revolucionando la comunicación entre aplicaciones o, de forma más general, la comunicación entre equipos, pues ofrece un formato de datos universal que permite adaptar o transformar fácilmente la información.

Para nuestro sistema, este lenguaje es una herramienta fundamental, pues lo utilizamos para crear el fichero de configuración, mediante el cual el módulo será capaz de hacer las búsquedas en Internet. A través de las etiquetas en las que se apoya este metalenguaje, se pueden distribuir y organizar los datos precisos para que el robot realice las descargas con mayor facilidad.

1.5 Servicio Windows

Los servicios de Microsoft Windows, antes conocidos como servicios NT, permiten crear aplicaciones con un tiempo de ejecución largo que corren en sus propias sesiones de Windows. Estos servicios pueden iniciarse automáticamente cuando se inicia el sistema, se pueden pausar y reiniciar, y no muestran ninguna interfaz de usuario. Esto hace que los servicios resulten perfectos para ejecutarse en un servidor o allí donde se necesite una funcionalidad de ejecución larga que no interfiera con los demás usuarios que trabajen en el mismo equipo. También puede ejecutar servicios en el contexto de seguridad de una cuenta de usuario específica, diferente de la del usuario que inició la sesión o de la cuenta predeterminada del equipo.

Un Servicio de Windows no es una aplicación normal, ya que no tiene interfaz gráfica de cara al usuario y tampoco es una especie de Servicio Web que funciona en Windows en lugar de hacerlo en un sitio de Internet. Los Servicios de Windows son aplicaciones que funcionan sin interactuar directamente con el usuario y por regla general se inician junto con el sistema, sin que ningún usuario tenga que iniciarlo; pues son programas o aplicaciones cargadas por el propio sistema operativo. Estas aplicaciones tienen la particularidad que se encuentran corriendo en segundo plano (Background).

Los servicios pueden encontrarse en dos estados posibles. Pueden estar iniciados, es decir, se encuentra ejecutándose/corriendo o puede estar detenido. Además de puede pausar o reanudar un servicio utilizando el Administrador de control de servicios, desde el Explorador de servidores, o por medio de llamadas a métodos en el código.

Y tienen tres opciones posibles de inicio:

- Automático: Se inician junto con el sistema operativo.

- Manual: Podemos iniciarlo y detenerlo manualmente cuando queramos u otro servicio puede hacerlo automáticamente. En un principio estaría detenido.

- Deshabilitado: No se puede iniciar manualmente ni otro servicio puede hacerlo.

1.6 Expresiones regulares

El Lenguaje Universal es demasiado amplio y no permite ningún tipo de restricciones a la hora de definir las secuencias de caracteres. Por eso nos interesa definir lenguajes más restrictivos que nos permitan localizar solamente aquellas cadenas de texto (secuencias de caracteres del alfabeto) que nos interesan.

Una Expresión Regular es un patrón que describe a una cadena de caracteres, siendo la clave para conseguir un procesamiento de texto potente, flexible y eficiente. Pues nos sirve para definir lenguajes, imponiendo restricciones sobre secuencias de caracteres, que permiten describir y parsear texto mediante una notación de patrones propia. Por tanto una Expresión Regular estará formada por el conjunto de caracteres del alfabeto original, más un pequeño conjunto de caracteres extra (meta-caracteres), que nos permitirán definir estas restricciones, pues este tipo de caracter básico es el que proporciona a las expresiones regulares su eficacia de procesamiento.

Las expresiones regulares se utilizan para hacer búsquedas contextuales y modificaciones sobre textos. A pesar de que estas están muy extendidas por el mundo de Unix, no existe un lenguaje estándar de expresiones regulares. Más bien se puede hablar de diferentes dialectos.

Existen, por ejemplo, representantes del conocido programa grep, egrep y fgrep. Estos usan expresiones regulares con capacidades ligeramente diferentes. Perl se puede calificar como el lenguaje con la sintaxis de expresiones regulares más desarrollado, aunque también existen el Sed y el Awk. Por suerte todos estos dialectos siguen los mismos principios y en el momento que se han entendido, el resto es sencillo.

Las expresiones regulares proporcionan un método eficaz y flexible para procesar texto. La notación extensiva de búsqueda de patrones coincidentes de

las expresiones regulares le permite analizar con rapidez grandes cantidades de texto para buscar patrones de caracteres específicos, para extraer, modificar, reemplazar o eliminar subcadenas de texto o para agregar las cadenas extraídas a una colección con objeto de generar un informe. Para muchas de las aplicaciones que manejan cadenas (como el procesamiento HTML, el análisis de archivos del Registro y el análisis de encabezados HTTP), las expresiones regulares constituyen una herramienta indispensable.

Las expresiones regulares que utilizará nuestro sistema son las usadas en Microsoft .NET Framework, debido a que esta es la plataforma en la que se desarrollará la aplicación; en él se incorporan las funciones más comunes de otras implementaciones de expresiones regulares, como las de Perl y awk.

Las clases de expresiones regulares de .NET Framework forman parte de la biblioteca de clases base y pueden utilizarse con cualquier lenguaje o herramienta que trabaje con Common Language Runtime, como ASP.NET y Visual Studio .NET.

1.7 Herramientas utilizadas en la aplicación

1.7.1 Lenguajes y tecnología

1.7.1.1 Plataforma .NET, C# como lenguaje de programación

La plataforma .NET es un nuevo entorno de programación especialmente diseñado para la creación de aplicaciones y de servicios web. Podría decirse que supone un cambio tan grande en el modo de programar como en su día lo fue la transición desde MS-DOS a Windows. Ahora el programador no trabaja directamente contra un sistema operativo concreto como Windows, sino que lo hace frente a una máquina virtual (el CLR o Common Language Runtime) que le ofrece los servicios que antes le proporcionaba el sistema operativo de forma más simplificada y adecuada a los tiempos actuales.

La mejor forma de resumir de las características de la plataforma .NET es enumerar los servicios que proporciona el CLR a todas las aplicaciones que desarrolladas para la misma. Entre éstas destacan las siguientes:

- Sencillo modelo de programación.

- Tratamiento homogéneo de errores mediante excepciones
- Desarrollo interlenguaje
- Ejecución multiplataforma
- Gestión automática de memoria con recolección de basura
- Aislamiento de procesos
- Soporte multihilo
- Seguridad avanzada basada en el usuario y la procedencia del código
- Interoperabilidad con código antiguo

Una de las principales y más novedosas características de la plataforma .NET que acaba de señalarse es su orientación hacia el desarrollo interlenguaje. Esto significa que en un proyecto .NET puede escribirse cada clase en cualquiera de los diferentes lenguajes que se han adaptado para funcionar en .NET, y la integración entre las clases escritas en los diversos lenguajes se hará perfecta y transparentemente, siendo incluso posible definir en cualquier lenguaje clases derivadas de clases escritas en cualquiera otro lenguaje.

Se han adaptado a .NET la inmensa mayoría de lenguajes de programación existentes, como C++, JScript, Visual Basic, Java, Eiffel, Cobol, Perl, Python, Fortran, Pascal, Smalltalk, etc. Sin embargo, Microsoft también ha creado un nuevo lenguaje llamado C# especialmente recomendado para la programación de las aplicaciones .NET y al que se suele bautizar como el lenguaje estrella de .NET.

Es un lenguaje totalmente orientado a objetos y además tipado que facilita la detección de errores e impide por defecto la creación de software altamente inseguro, pero también ofrece ciertas facilidades para la creación de trozos de código delicados por ejemplo utilizando punteros. De esta forma se consigue un equilibrio entre seguridad y flexibilidad.

1.7.1.2 SQL –Server 2000

SQL o Lenguaje Estructurado de Consultas, soporta al motor de base de datos cliente/servidor SQL Server, el cual está diseñado para almacenar datos en un sitio central llamado servidor (pueden ser varios) y distribuirlos a otros sistemas

llamados clientes. Éstos realizan consultas al servidor, el cual los procesa y, luego, entrega los resultados (conjunto de registros) a los clientes que los solicitaron. La ventaja de esta arquitectura es que sus requerimientos de hardware no son demasiado exigentes, aunque sí es conveniente poseer un equipamiento robusto del lado del servidor.

1.7.1.3 ADO.NET

El acceso a bases de datos es una constante en muchísimas aplicaciones. ADO son las siglas de ActiveX Data Objects, una de las tecnologías de Microsoft para comunicar programas con bases de datos. ADO.NET es una evolución del modelo de acceso a datos de ADO que controla directamente los requisitos del usuario para programar aplicaciones escalables. Se diseñó específicamente para el Web, teniendo en cuenta la escalabilidad, la independencia y el estándar XML.

ADO.NET proporciona acceso a un amplio abanico de bases de datos, entre ellas a MySQL, ODBC, Oracle, OLE DB, Microsoft SQL Server, entre otras.

Además incorpora una característica interesante y es que permite realizar consultas y modificaciones a una base de datos de forma off-line, es decir:

1) Se realiza una conexión inicial para traer los datos necesarios

2) Se libera la conexión

3) A continuación se trabaja con esos datos, como hemos liberado la conexión estamos permitiendo que otros clientes se puedan conectar a la base de datos.

4) Y en caso de que sea necesario se realiza una nueva conexión para realizar las modificaciones en el servidor.

1.7.2 Metodología utilizada para modelar

Racional Unified Process (RUP)

Un proceso define *Quién* está haciendo *Qué*, *Cuándo* y *Cómo* para lograr cierto objetivo. En la ingeniería de software el objetivo es construir un producto de software ó mejorar alguno existente.

RUP como nueva metodología para modelar procesos, captura varias de las mejores prácticas en el desarrollo moderno de software en una forma que es aplicable para un amplio rango de proyectos y organizaciones, siendo una guía de cómo utilizar de manera efectiva UML; provee a cada miembro de un equipo de trabajo un fácil acceso a una base de conocimiento con guías, plantillas y herramientas para todas las actividades críticas de desarrollo; y crea y mantiene modelos, en lugar de enfocarse en la producción de una gran cantidad de papeles de documentación. Tiene varias fases: *Inicio*, donde se define el alcance del proyecto, *Elaboración*, donde se crea el plan del proyecto, se especifican las características y la arquitectura base, *Construcción*, donde se construye el producto y la ultima etapa es la de *Transición*, en la que se realiza la transición del producto a la comunidad del usuario.[8]

Algunas de las mejores prácticas que RUP emplea son:

- Incrementa la productividad del equipo:

  pues todos los miembros comparten una base de conocimiento, un proceso, una misma vista de cómo desarrollar software y un mismo lenguaje de modelación (UML).

- Posibilita el desarrollo interactivo del software:

  permitiendo un entendimiento incremental del problema a través de refinamientos sucesivos, nos brinda metas específicas que permiten que el equipo de desarrollo mantenga su atención en producir resultados, midiendo el progreso según avanzan las implementaciones.

- Brinda una modelación visual del Software

  a través de la cual se captura la estructura y comportamiento de arquitecturas y componentes, se muestra como encajan de forma conjunta los elementos del sistema, se mantiene la consistencia entre un diseño y su implementación y se promueve una comunicación no ambigua entre sus componentes.

- Verifica la calidad del software

  Pues crea pruebas para cada escenario (casos de uso) para asegurar que todos los requerimientos están propiamente implementados y verifica la calidad del software con respecto a los requerimientos

basados en la confiabilidad, funcionalidad, desempeño de la aplicación y del sistema.

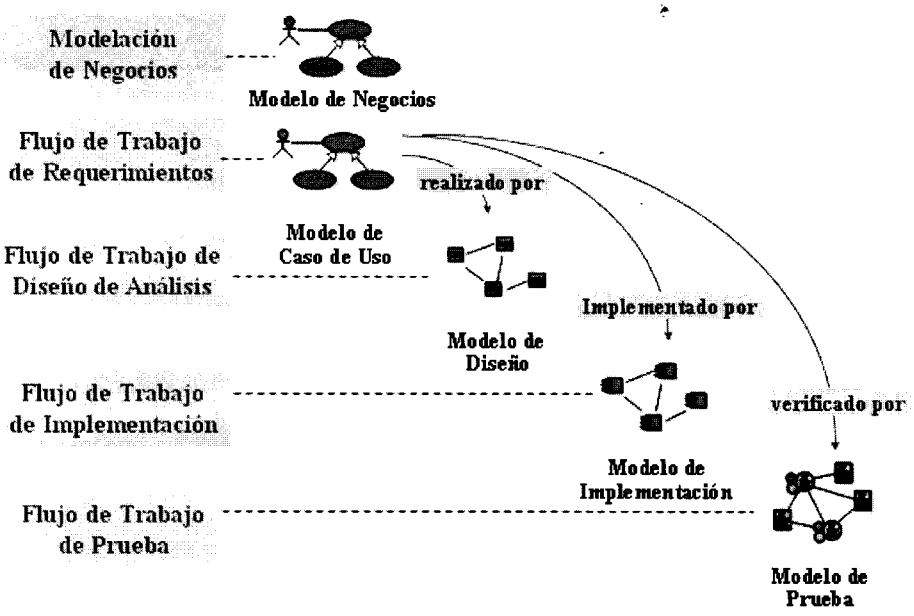


Fig 1.1 RUP. Modelos y flujos de trabajo

1.8 Conclusiones

En este capítulo hemos tratado de ubicar al lector de forma general en el medio en que se ha desarrollado nuestro sistema. Intentando introducirlo en el mundo de los proveedores automáticos de contenido, específicamente en aquellos que alimentan una base de datos de noticias actualizadas.

Se ha hecho un pequeño recorrido por los más conocidos proveedores automáticos de noticias y se han mostrado las facilidades y utilidades que brindan algunos de ellos, destacando el Google News, por ser uno de los motores de búsqueda más completos.

De la misma forma se ha profundizado en el estudio de los web crawlers, por considerarse como el patrón de búsqueda a seguir para desarrollar nuestra aplicación. Haciendo referencia además al resto de las herramientas utilizadas.

Capítulo 2.

Estudio Preliminar

2.1 Introducción

En este capítulo se da una idea general del problema al que nos enfrentamos: el sitio de la intranet universitaria no cuenta con un sistema que lo provea de noticias actualizadas de forma automática, además se profundiza en la solución que proponemos. Se describen los objetos de estudio y automatización, los principales problemas existentes en cuanto a la prestación de servicios noticiosos en la Intranet de la Universidad de las Ciencias Informáticas, se analizan los requerimientos funcionales y no funcionales, y los casos de uso del sistema.

2.2 Objeto de Estudio

Una necesidad es la carencia de algo, que en ocasiones experimenta una persona. Esa necesidad puede ser reconocida o no por el sujeto; sin embargo, cuando ocurre lo primero, se espera que sea satisfecha.

En consecuencia, la necesidad de información se asume como el posible reconocimiento de una carencia o la aceptación de un estado anómalo del conocimiento por parte del usuario bien sea por sí mismo o porque alguien le ayudó a reconocerla, aunque en muchas ocasiones aquel no sepa expresar esa carencia.

Uno de los sectores más sensibles en el mundo es el educativo, por la influencia que ejercen los jóvenes en la sociedad en la que están insertados, de ahí la necesidad de que los universitarios tengan a su alcance herramientas que les permitan mantenerse informados con los últimos sucesos ocurridos. Este proceso de formación de un nivel informativo sólido depende, en gran medida, del nivel de acceso a la información. Evitar que llegue a nuestros jóvenes información tergiversada es de vital importancia, pues "Todo hombre o mujer vive una cuenta regresiva. Hace mucho tiempo que hemos entregado a nuestra causa cada minuto de vida que nos reste."[9]

Por ello es necesario estudiar y desarrollar diferentes formas de obtener, almacenar y brindar lo más detalladamente posible cada recurso que pueda mantenernos actualizados, para convertirnos en las personas preparadas que estos nuevos tiempos precisan.

2.2.1. Situación problémica

Podemos decir que entre los fenómenos adversos que trae consigo la realización manual de este trabajo y que frenan la posibilidad de ofrecer noticias actualizadas tenemos:

- No existe personal dedicado solo a la realización de esta tarea, quien la realiza debe hacerlo como un apéndice más a su faena diaria.
- El tiempo que esta persona debe dedicar diariamente a visitar sitios noticiosos en Internet y seleccionar de ellos las noticias que pueden ser publicadas en algún momento.
- Descargar y almacenar en una base de datos, noticias diariamente para publicarlas luego.
- Pueden existir publicadas en la intranet noticias que no hayan ocurrido en un plazo de menos uno o dos días aproximadamente.

2.2.2. Problema

Este software, al formar parte del programa de informatización de la Universidad de las Ciencias Informáticas, "UCI Ciudad Digital", y con el proceso de modernización que este trae consigo, persigue brindar un servicio automatizado que mediante un funcionamiento estable, mantenga actualizadas las noticias de la intranet universitaria.

Quizás una posible solución para ello sería contratar personal que se dedique sólo a la realización de esta tarea, es decir, a la búsqueda, descarga y almacenamiento de las noticias diariamente, pero no estaríamos siendo eficientes. Con la terminación de este producto no necesitaremos presencia humana, solo unos pocos minutos de una persona, que será la encargada de dar mantenimiento al sistema y asegurarse de que este no dejará de funcionar,

aún cuando ocurran cambios en la distribución de las páginas de los sitios que se visitan, pues mantendrá actualizado el fichero de configuración que funge como médula espinal de la aplicación.

2.2.3 Ubicación.

Con el objetivo de crear un centro de altos estudios de nuevo tipo, donde los estudiantes sean capaces de vincular el binomio estudio-producción, surge la Universidad de las Ciencias Informáticas (UCI) siendo la universidad que rompe con los esquemas y paradigmas de las más longevas universidades cubanas. La formación docente-productiva de sus alumnos debe contribuir en gran medida al desarrollo de la ingeniería informática en nuestro país, y además ser puntal en la industria del software cubano para el trabajo de informatización de la sociedad.

Como parte de la DIP de Soporte de Software, que a la vez es una de las que forman el Grupo de Informatización, encargado de digitalizar la mayor cantidad de los servicios que presta la Universidad, dentro del proyecto "UCI, Ciudad Digital", este software facilitará el proceso de descarga, almacenamiento y publicación de noticias en la intranet universitaria.

Este producto no solo puede ser utilizado en la UCI, todas las empresas y entidades con acceso a Internet y un sitio donde realizar las publicaciones pueden auxiliarse de él, en aras de mantener actualizado su espacio noticioso.

2.3 Objeto de automatización

La Intranet no cuenta con un sistema que de manera automática brinde las más recientes noticias ocurridas tanto en el ámbito nacional como internacional, dificultando el trabajo diario de las personas que diariamente deban dedicarse a realizar la actualización noticiosa de la Intranet Universitaria.

Esta aplicación tiene como objetivo proveer a la UCI de un módulo capaz de realizar las labores de búsqueda y descarga diarias de noticias de determinados sitios de Internet, su clasificación y almacenamiento para una posterior publicación.

Para realizar estas acciones, utilizaremos un servicio que irá obteniendo al leer de un fichero XML, la ubicación precisa de las noticias que quiero obtener en los sitios de Internet que ya han sido analizados y estudiados con anterioridad, para realizar la descarga de las mismas, así como su clasificación, luego serán almacenadas en una base de datos, de la cual podrán ser extraídas y seleccionadas las que deban publicarse en la Intranet. Para ello contamos con:

- Fichero de configuración XML contiene la ubicación exacta de la información que deseamos obtener, con el objetivo de poder realizar en él los cambios necesarios en caso de que en las páginas de los diversos sitios de Internet varíen su disposición.
- Servicio de Windows encargado de despertar y realizar la lectura del fichero de configuración para la descarga de la información.
- Módulo para el procesamiento y análisis de la información extraída, así como su clasificación.
- Módulo para el trabajo con la Base de Datos, donde estarán insertadas las noticias ya clasificadas conjuntamente con las imágenes que dicha noticia pueda tener.

2.4 Información que se maneja

El sistema tiene como médula central para lograr cierta flexibilidad, un fichero de configuración, a través del cual el módulo obtiene toda la información necesaria con un nivel profundo y detallado de la ubicación de cada noticia en Internet, que luego será descargada, procesada y almacenada en la base de datos. Veamos una breve descripción de este fichero, para ver un fragmento del mismo diríjase al **Anexo 1**:

Fichero de configuración

El fichero está escrito en formato XML y está dividido en dos segmentos principales, el primero con la información necesaria para extraer las noticias de las fuentes y el segundo con la información que utilizará en Servicio de Windows en su ejecución.

En lo relativo al primer segmento, la estructura del fichero de configuración se divide en:

<conf>	Etiqueta global del fichero, contiene el resto de las etiquetas
<site>	Entre estas etiquetas se encuentra toda la información relacionada con una fuente de noticias, hasta el momento solo contamos con dos sitios, Granma y CNN en español.
<url>	Contiene la dirección de la fuente de noticias en Internet.
<step>	El procesamiento de una fuente noticiosa está compuesto por un conjunto de pasos. Entre estas etiquetas se encuentran todas las operaciones a realizar en cada paso.
<operation>	Tiene dos atributos, "tag" y "cant". Se utilizan para localizar el fragmento del código HTML que se quiere procesar.

El segundo segmento se divide en:

<ontime>	Tiene un atributo "time" que indica la hora, en formato militar, en que el Servicio Windows iniciará el proceso.

Las noticias serán almacenadas en una base de datos de donde podrán ser extraídas, tras previa revisión del personal adecuado, para su posterior publicación. Esta base de datos, llamada PA_Noticias, está formada por dos tablas: Not_Textos y Not_Img. La primera almacena todo el texto de la noticia, con varios campos en los que se recoge toda la información que necesitamos de las noticias: la fecha de publicación, la clasificación y la categoría, donde se señala si la noticia hace referencia al acontecer mundial o al de nuestro país, si es deportiva, cultural, de carácter científico técnica, etc. y la fuente de donde se obtuvo. Por su parte, la tabla Not_Img contiene las imágenes que puede tener una noticia.

2.5 Propuesta de sistema

Analicemos detalladamente las propuestas que ofrece nuestro sistema para resolver el problema existente en la Universidad de las Ciencias Informáticas con respecto a la actualización manual de las noticias de la Intranet a través de su funcionamiento.

Al ser un servicio de Windows la solución que brindamos, ha de encontrarse corriendo en background, hasta el momento en que el sistema operativo lo inicie, la hora de comienzo de la aplicación forma parte de la información contenida en un fichero XML, llamado fichero de configuración, de donde el servicio lee constantemente. Luego de iniciarse, el servicio continúa leyendo de este fichero pues en él se encuentran todos los datos precisos de la ubicación exacta de los artículos que serán descargados de Internet. Utilizando expresiones regulares se van macheando las páginas hasta ubicar el fragmento con la información, cuyo formato es HTML, por lo que debe pasar por un proceso de purificación antes de ser almacenada en la base de datos, para tener solo en ella el texto íntegro de la noticia sin etiquetas o posibles comentarios que el documento traiga consigo y las posibles imágenes que el artículo posea; todo este proceso se realiza utilizando funciones auxiliares, que devolverán al culminar solo el texto con la noticia ya clasificada, categorizada, con fecha de publicación y además la ubicación de sus imágenes.

Si comparamos la propuesta de nuestro sistema con el proveedor de noticias de Google, GoogleNews, que es un sistema de renombre mundial, hallaremos muchas diferencias. Nuestra aplicación tiene como principal diferencia la presencia del fichero de configuración XML, cuya ventaja está en tener la posibilidad de hacer cambios en él si se produjeran variaciones en la configuración de las páginas, sin tener que alterar el código del módulo. Además nos da la posibilidad de añadir o eliminar sitios, en fin, hacer variaciones no inherentes al código del sistema. Por su parte tenemos varias desventajas, entre ellas el no poder mostrar las noticias en su formato original, no tener acceso a todos los hipervínculos que puedan permitirnos tener el texto de una noticia, no tener titulares, etcétera. Aunque la comparación no deba

establecerse entre dos sistemas tan diferentes, hemos de tenerlo en cuenta como paradigma de nuestra aplicación.

2.6 Especificación de Requerimientos del software.

2.6.1 Requerimientos funcionales

Nuestro sistema debe responsabilizarse con un grupo de acciones, que serán denominadas requisitos funcionales, ellas son:

1. Crear expresiones regulares para el trabajo de descarga del sistema.
2. Extraer el contenido de las páginas
   2.1 Diseñar algoritmos para la extracción de la información utilizando expresiones regulares.
3. Analizar el contenido extraído
   3.1 Limpiar el texto, seleccionando sólo el contenido que necesitamos a través de expresiones regulares.
4. Diseñar e implementar una base de datos para almacenar toda la información obtenida, incluyendo las imágenes.
   4.1 Utilizando expresiones regulares, determinar si en el contenido que estamos analizando hay presencia de imágenes.
5. Crear de un Servicio de Windows que se encargue de leer del fichero de configuración, obtenga el contenido deseado y las inserte después de haber sido analizadas, en la base de datos.

2.6.2 Requisitos no funcionales

Las características que debe tener el producto que obtengamos para que su funcionamiento sea óptimo, así como para asegurar su confiabilidad y seguridad, las llamamos requisitos generales o no funcionales. En nuestro sistema ellos son:

Requisitos de funcionalidad

1. Tiempo que deben utilizar los usuarios encargados de seleccionar la posible información a publicar y el administrador del sistema para entrenarse en la utilización de la aplicación

   1.1 Debe planificarse un tiempo de entrenamiento de una semana a lo sumo para el administrador.

Requisitos de confiabilidad

1. Disponibilidad

   1.1 El administrador decide la hora en que desee se realice la descarga, luego la información estará disponible en la base de datos por el tiempo que él decida.

2. Precisión

La precisión de este sistema depende de que en el fichero de configuración esté la ubicación exacta de toda la información que necesitamos extraer de Internet, para ello:

   2.1 Se debe tener conocimiento pleno de la estructura de las páginas a analizar para luego:

      2.1.1 Delimitar la información específica que se desee extraer.

      2.1.2 Definir la ubicación exacta de esta información.

   2.2 Para crear el fichero de configuración debemos

      2.2.1 Definir una etiqueta raíz que contenga el resto de las etiquetas.

      2.2.2 Definir etiquetas para cada página a analizar.

      2.2.3 Definir etiquetas para la ubicación exacta de la información a extraer dentro de cada página, clasificadas por categorías.

Requisitos de diseño e implementación.

1. Herramientas de desarrollo gráfico.

   1.1. Para realizar la modelación del sistema se utiliza como herramienta de desarrollo gráfico el Rational Rose.

2. Herramientas para el desarrollo de la aplicación.

2.1. Como lenguaje de programación se utiliza el C# por las facilidades que brinda para este tipo de aplicaciones.

3. Herramientas para el almacenamiento de la información

3.1. Como sistema gestor de la base de datos utilizamos Microsoft SQL Server 2000

Requisitos de software y de hardware.

Para lograr un buen funcionamiento de este sistema necesitamos una máquina con conexión a Internet y al menos instalada la versión cliente de la aplicación SQL Server.

Requerimientos legales.

Esta aplicación forma parte del conjunto de software perteneciente a la DIP de Soporte de Software, dentro del equipo de informatización que desarrolla el Proyecto UCI Ciudad Digital.

2.7 Definición de los casos de uso

2.7.1 Definición de los actores

Un actor de un sistema es una entidad externa del él, que de alguna forma participa en el caso de uso. Generalmente estimula al sistema con eventos de entrada o recibe algo de él.

Nombre del actor	Descripción
Sistema Operativo	Es un actor abstracto. Es el encargado de iniciar el servicio que se encontrará corriendo en background.

Tabla 2.1: Descripción del actor

2.7.2 Listado de casos de uso

Los casos de uso describen la funcionalidad del sistema. Es una herramienta para modelar el contexto de un sistema o para modelar los requisitos del mismo. [10]

CU-1	Iniciar el Servicio de Windows
Actor	Sistema Operativo (actor abstracto)
Descripción	El Servicio se encuentra corriendo en background leyendo del fichero de configuración. En él aparece especificada la hora en la que el módulo debe ser iniciado, luego el sistema operativo se encarga de iniciarlo y el módulo comienza la búsqueda, a medida que va obteniendo del fichero XML la ubicación de la información que debe hallar, luego inicia la descarga de todos los artículos y los va almacenando en una Base de Datos el texto de la noticia ya clasificada conjuntamente con las imágenes que pueda tener.
Referencia	RF – 5

Tabla 2.3 Descripción del Caso de Uso # 2

CU-2	Extraer Contenido
Actor	Sistema Operativo (actor abstracto)
Descripción	El Servicio obtiene del fichero XML la ubicación exacta de la información que debe hallar, luego utilizando expresiones regulares, machea las páginas de los sitios a visitar e inicia la descarga de todos los artículos tras haberlos encontrado.
Referencia	RF – 1, RF- 2

Tabla 2.4 Descripción del Caso de Uso # 2

CU-3	Analizar Contenido
Actor	Sistema Operativo (actor abstracto)
Descripción	Tras haberse descargado la página utilizando expresiones regulares, que han sido creadas con anterioridad, se realiza un proceso de limpieza del texto extraído, se eliminan los comentarios y los elementos del formato HTML distintos al texto de la noticia.
Referencia	RF – 1, RF – 3

Tabla 2.5 Descripción del Caso de Uso # 3

CU-4	Almacenar Contenido
Actor	Sistema Operativo (actor abstracto)
Descripción	Con toda la noticia presta para ser almacenada se insertan en la base de datos incluyendo, entre otros atributos, la clasificación en dependencia del origen de la fuente noticiosa, la categoría según el tema que aborda, la fecha, etc. De la misma forma se almacenan las imágenes que incluyan las noticias.
Referencia	RF - 4

Tabla 2.6 Descripción del Caso de Uso # 4

2.7.3 Diagrama de Casos de Uso del Sistema

Uno caso de uso es una secuencia de transacciones que son desarrolladas por un sistema en respuesta a un evento que inicia un actor sobre el propio sistema. Los diagramas de casos de uso sirven para especificar la funcionalidad y el comportamiento de un sistema mediante su interacción con los usuarios y/o otros sistemas. O lo que es igual, un diagrama que muestra la relación entre los actores y los casos de uso en un sistema. Una relación es una conexión entre los elementos del modelo. [11]

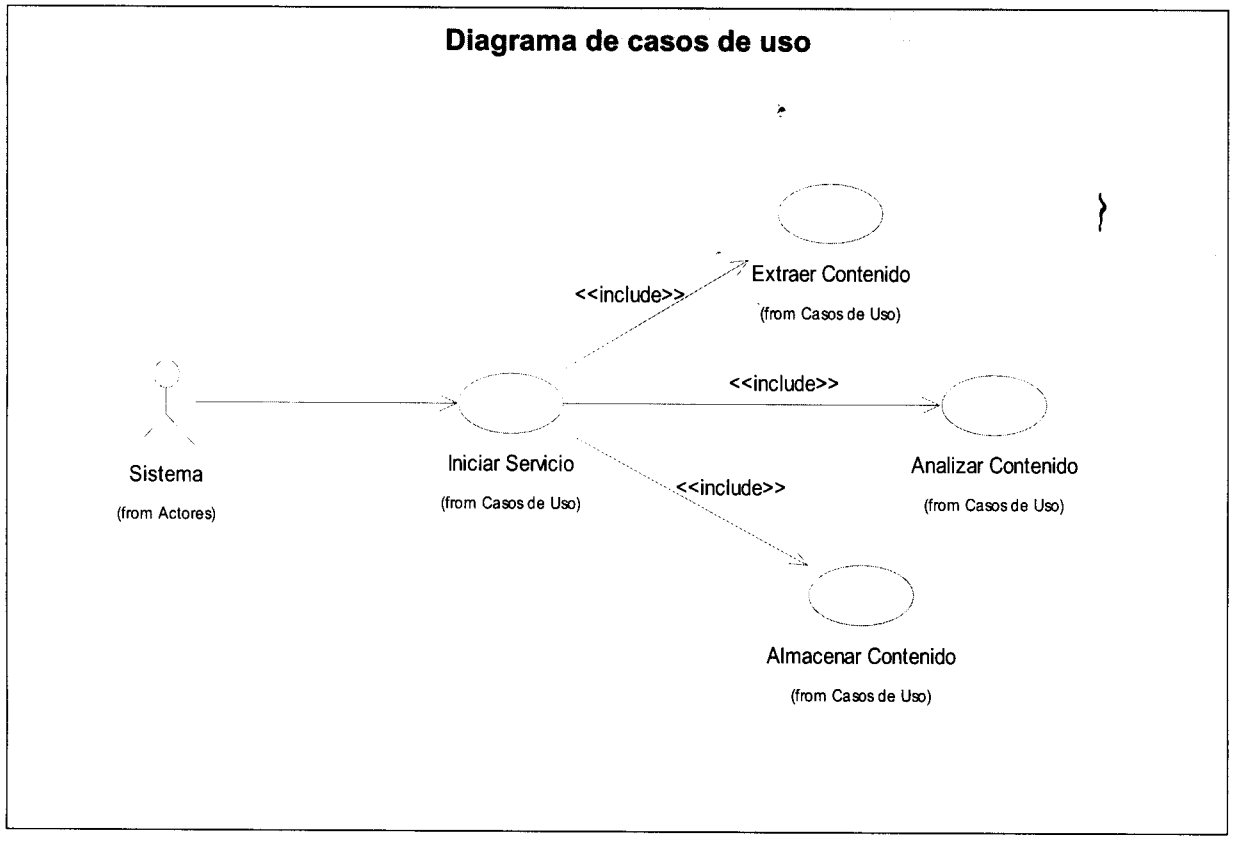


Fig. 2.1 Diagrama de Casos de Uso

2.7.4 Casos de uso por ciclo

Cód	Nombre de caso de uso	Paquete	Justificación de la selección.
CU-1	Iniciar el Servicio de Windows	CU Iniciar Servicio	Este caso de uso es quien contiene toda la etapa de funcionamiento de la aplicación. Incluye los procesos de lectura del fichero de configuración para realizar luego la descarga de los artículos de Internet, procesar después el texto y las imágenes que puedan traer las

			noticias y almacenarlas luego en una base de datos.
CU-2	Extraer Contenido	CU Extraer Contenido	Este caso de uso se encarga de todo el proceso de descarga de los artículos de Internet, a través del trabajo con expresiones regulares.
CU-3	Analizar Contenido	CU Analizar Contenido	Este caso de uso contiene todas las funciones que manejan el proceso de obtener las expresiones regulares y además en bruto el texto de la noticia, eliminando los posibles churres que pueda contener el texto en formato HTML.
CU-4	Almacenar Contenido	CU Almacenar Contenido	Este caso de uso se encarga de recolectar las noticias en la base de datos, ya clasificadas y categorizadas, conjuntamente con las imágenes que dicha noticia posea.

Tabla 2.7 Ciclo de desarrollo del software

2.7.5 Casos de uso expandidos

Dirigirse a los anexos del 2 al 6.

2.8 Conclusiones

En este capítulo hemos hecho referencia a las ideas a seguir para brindar un sistema que de manera automática provea de noticias actualizadas la intranet de la Universidad de las Ciencias Informáticas; evitando que este trabajo deba hacerse manual y contribuyendo con el Proyecto de informatización de la Universidad.

Se han detallado los procesos que forman parte del proceso de desarrollo de la aplicación. Haciendo énfasis a través los requerimientos del sistema y la explicación detallada de los casos de uso, de cada uno de los pasos a seguir para obtener un producto con una calidad garantizada.

Capítulo 3.

Análisis del Sistema.

3.1 Introducción

La esencia del los procesos de análisis y diseño del sistema consiste en situar el dominio de un problema y su solución lógica dentro de la perspectiva de los objetos.[12]

3.2 Análisis.

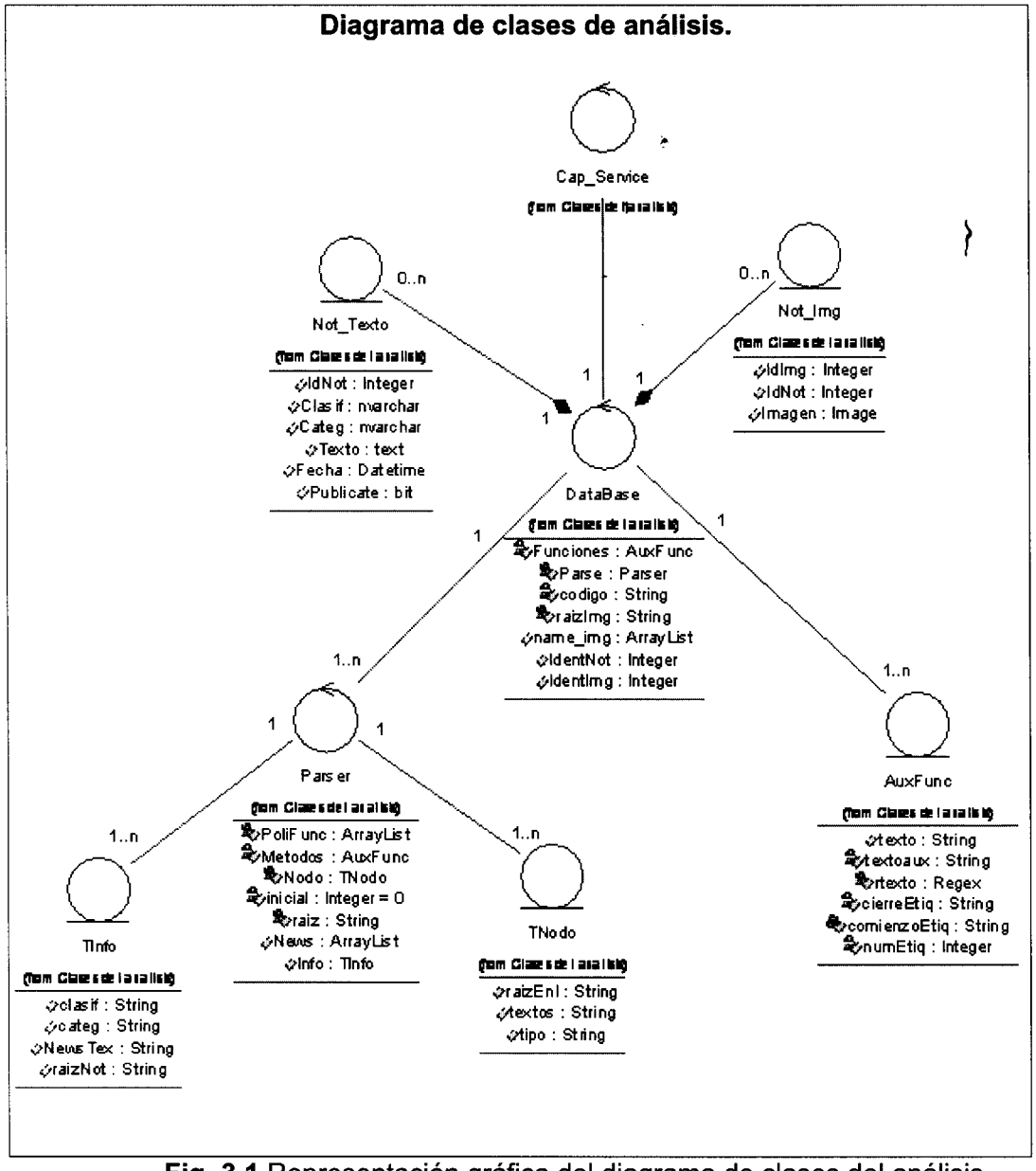


Fig. 3.1 Representación gráfica del diagrama de clases del análisis.

3.3 Diseño.

3.3.1 Diagramas de secuencia

3.2.2.1 Diagrama de secuencia Caso de Uso # 1

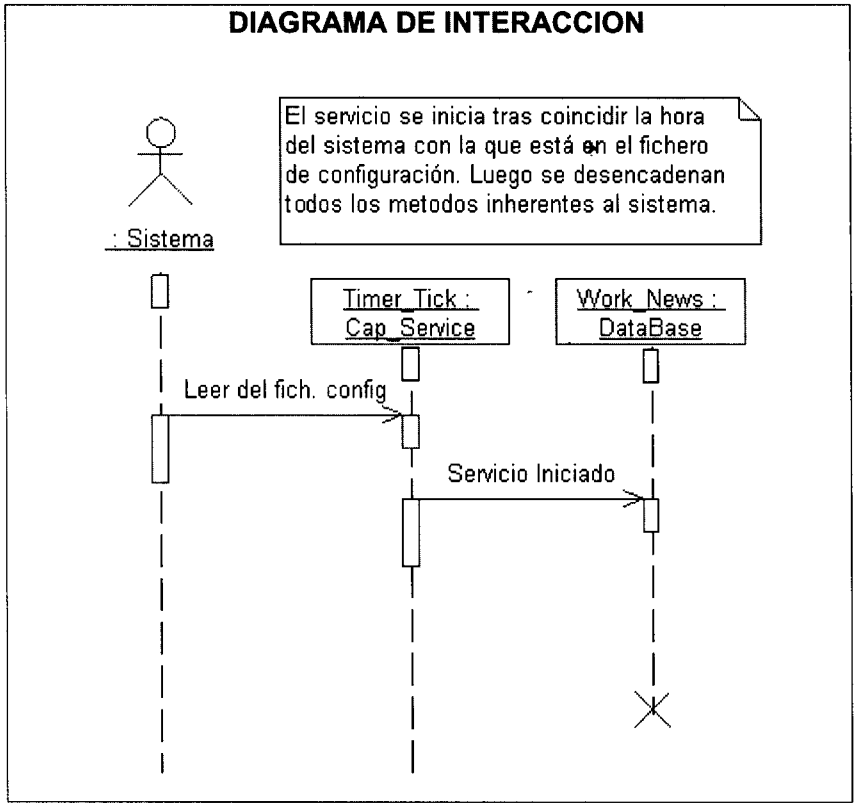


Fig. 3.3 Diagrama de secuencia CU # 1

3.2.2.2 Diagrama de secuencia Caso de Uso # 2

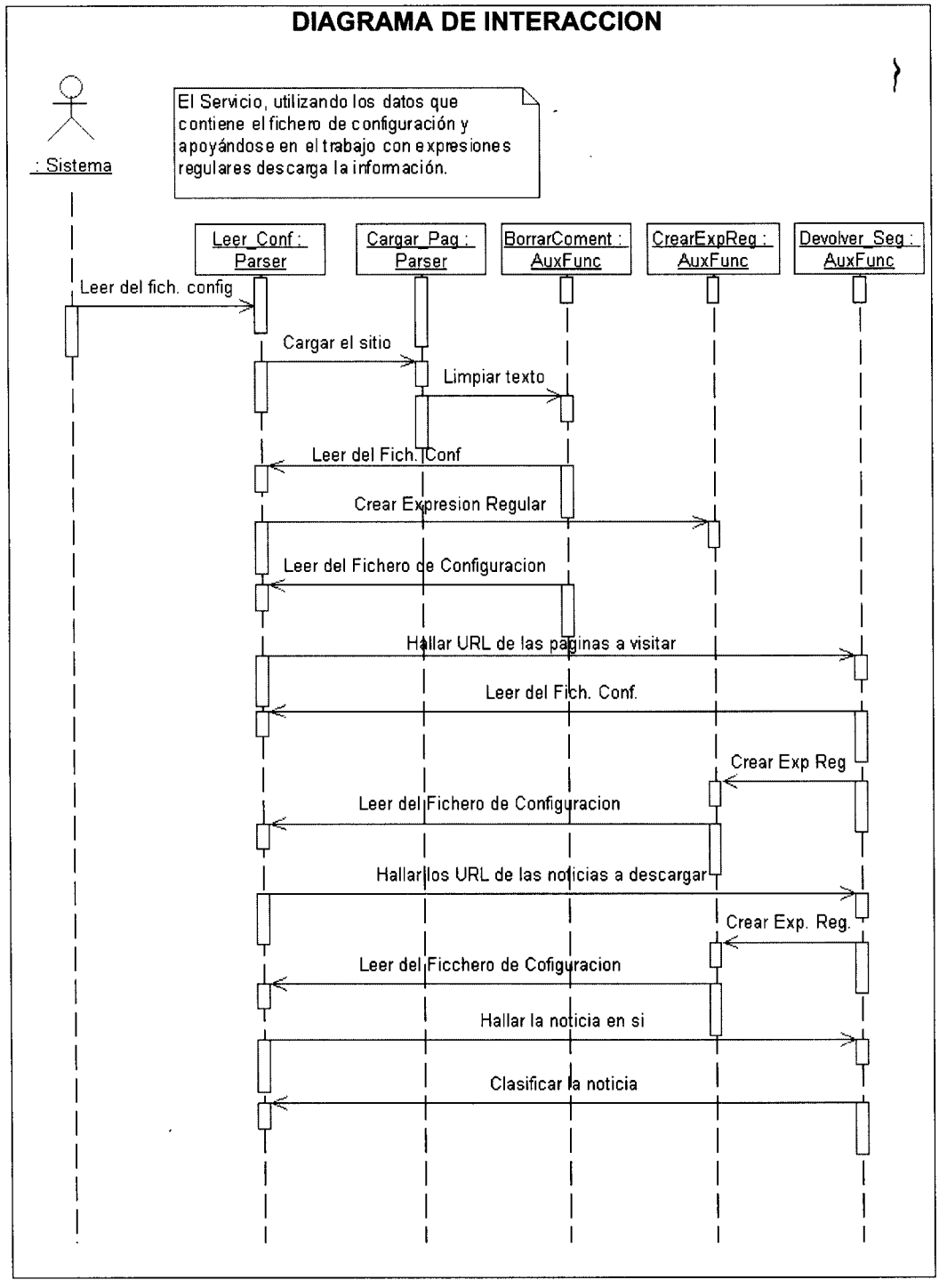


Fig. 3.4 Diagrama de secuencia CU # 2

3.2.2.3 Diagrama de secuencia Caso de Uso # 3

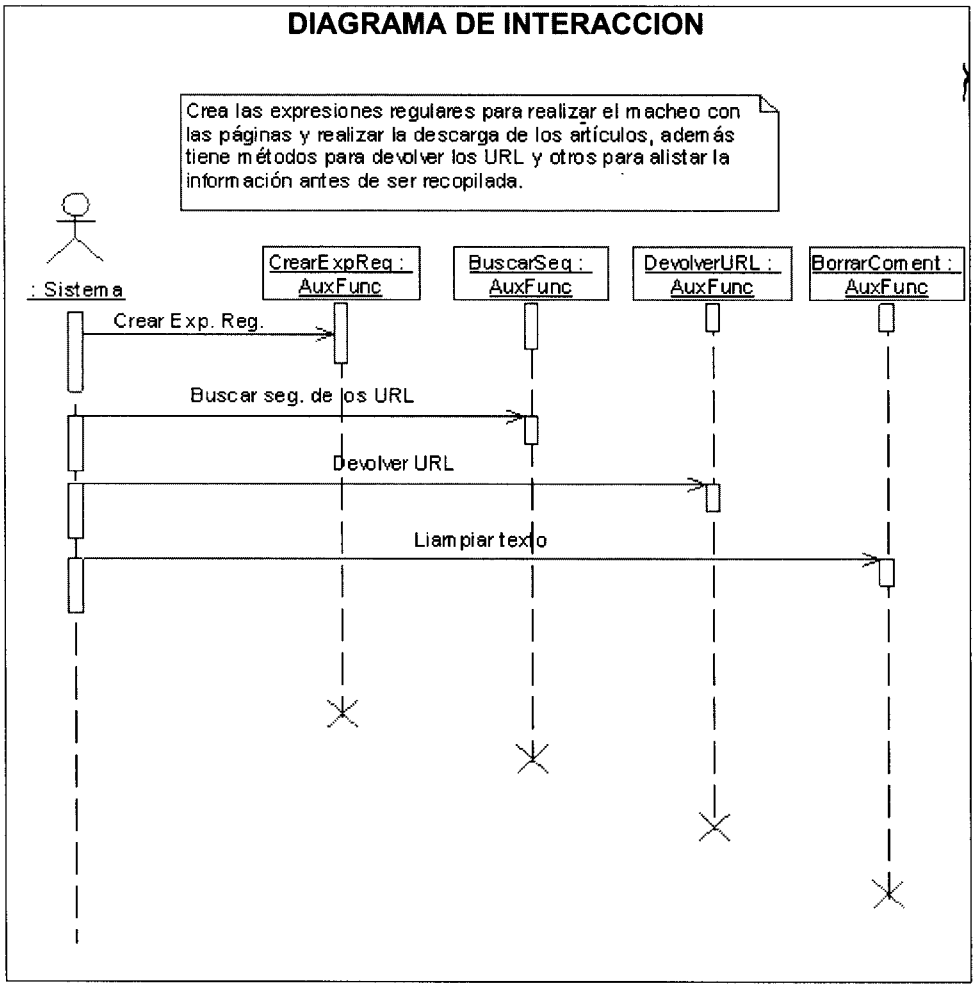


Fig. 3.5 Diagrama de secuencia CU # 3

3.2.2.4 Diagrama de secuencia Caso de Uso # 4

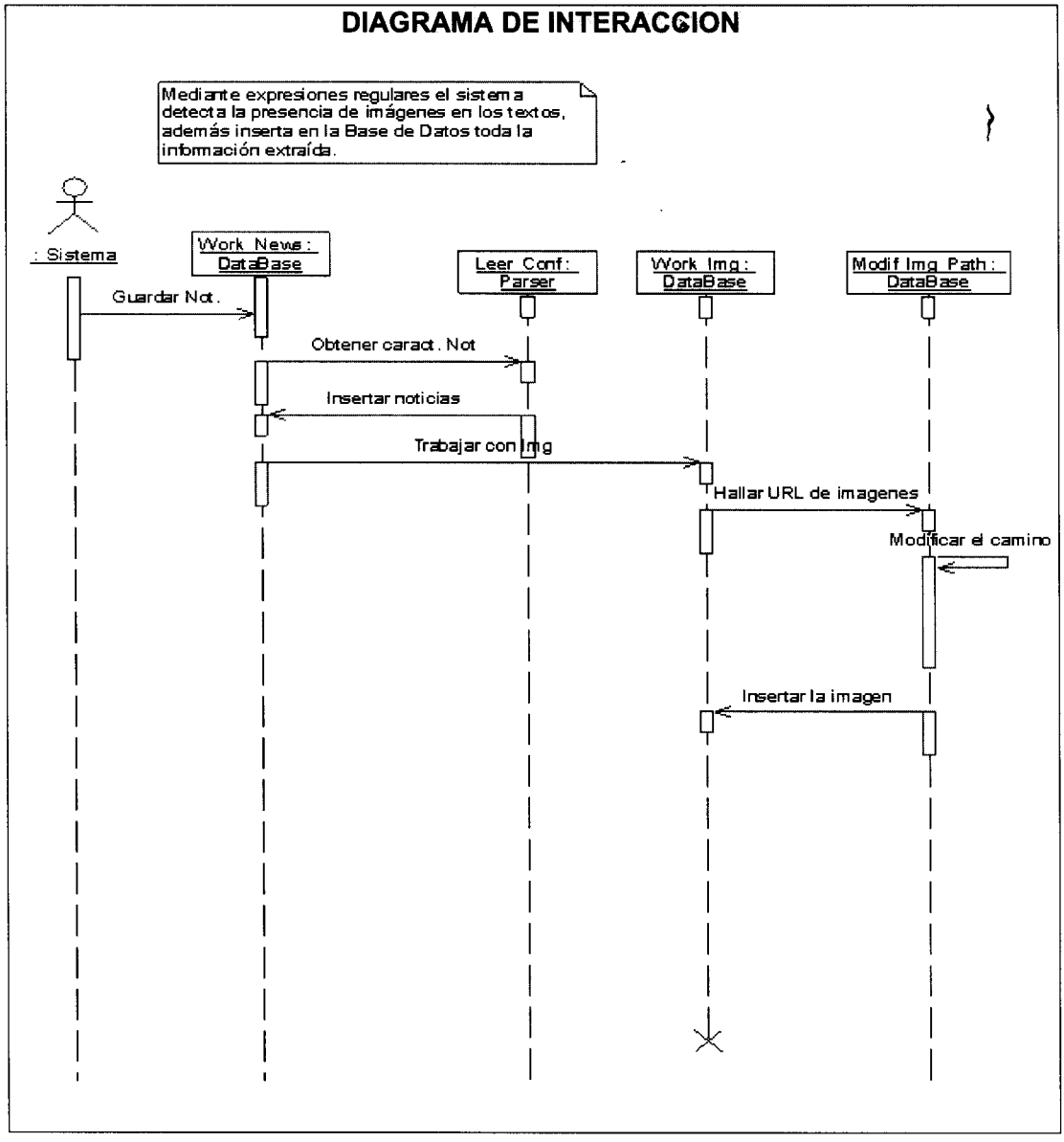


Fig. 3.6 Diagrama de secuencia CU # 4

3.2.2 Diagrama de clases del diseño

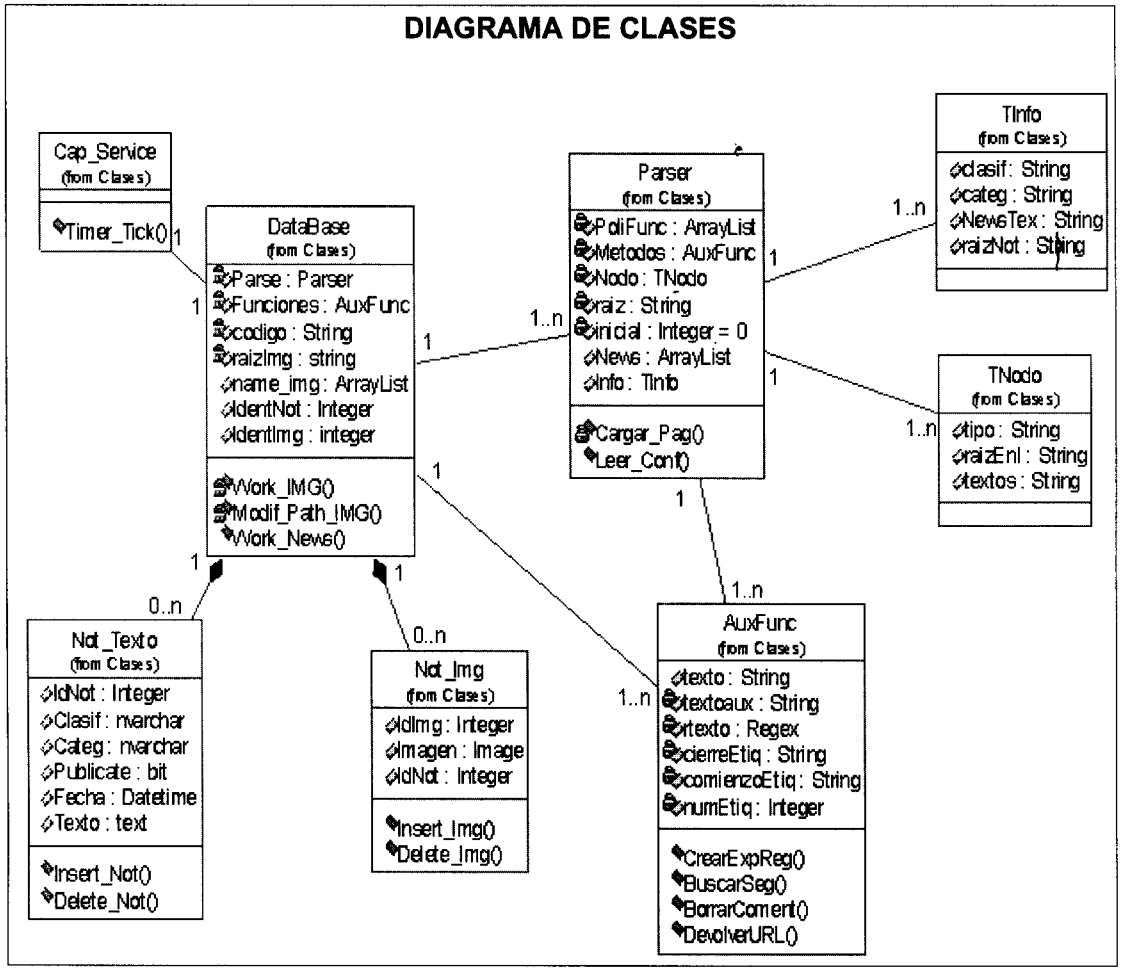


Fig. 3.7 Diagrama de clases del diseño

3.2.3 Descripción de las clases.

Nombre: Cap_Service	
Tipo de clase: controladora	
Atributo	**Tipo**
-	-
Para cada responsabilidad:	
Nombre:	Timer_Tick
Descripción:	Método encargado de iniciar el procesamiento automático de la búsqueda, descarga e inserción, haciendo una llamada al procedimiento Work_News de la clase DataBase cuando la hora del sistema coincida con la hora en que según el fichero de configuración debe iniciarse la

	aplicación.

Tabla 3.1 Descripción de la clase Cap_Service

Nombre: DataBase	
Tipo de clase: controladora	
Atributo	**Tipo**
Funciones	AuxFunc
name_img	ArrayList
raizImg	String
Parse	Parser
Codigo	String
IdNot	Integer
IdImg	Integer
Para cada responsabilidad:	
Nombre:	Work_News
Descripción:	Método encargado de dar inicio a la lectura del fichero de configuración, y tras haber obtenido todas las características de las noticias, asigna a cada atributo de la tabla Not_Texto su valor y hace la conexión con la base de datos para insertar esos valores. Además da inicio al trabajo con las imágenes.
Nombre:	Modif_Path_Img
Descripción:	Método encargado hallar mediante expresiones regulares, en el texto que se ha descargado la presencia de imágenes, luego modifica los enlaces de imágenes existentes en la noticia.
Nombre:	Work_Img
Descripción	Método que realiza el trabajo con las imágenes, actualiza los atributos de la tabla Not_Img que tiene las imágenes de cada noticia para ser insertadas en la base de datos.

Tabla 3.2 Descripción de la clase DataBase

Nombre: Not_Texto	
Tipo de clase: entidad	
Atributo	**Tipo**
IdNot	Integer
Clasif	Nvarchar
Categ	Nvarchar
Publicate	Bit
Fecha	DateTime
Texto	Text
Para cada responsabilidad:	
Nombre:	Insert_Not
Descripción:	Inserta las noticias conjuntamente con todas sus características, en la tabla Not_Texto.
Nombre:	Delete_Not
Descripción:	Elimina las noticias de la tabla Not_Texto.
Nombre:	Update_Not
Descripción:	Actualizar las noticias de la tabla Not_Texto

Tabla 3.3 Descripción de la clase Not_Texto

Nombre: Not_Img	
Tipo de clase: entidad	
Atributo	**Tipo**
IdNot	Integer
IdImg	Integer
Img	Image
Para cada responsabilidad:	
Nombre:	Insert_Img
Descripción:	Inserta las imágenes de cada noticia en la tabla Not_Img.
Nombre:	Delete_Img
Descripción:	Elimina las imágenes de cada noticia en la tabla Not_Img.
Nombre:	Update_Img
Descripción:	Actualizar las imágenes de cada noticia en la tabla

	Not_Img.

Tabla 3.4 Descripción de la clase Not_Img

Nombre: AuxFunc	
Tipo de clase: entidad	
Atributo	**Tipo**
comienzoEtiq	String
cierreEtiq	String
numEtiq	Integer
Textoaux	String
Rtexto	Regex
Texto	String
Para cada responsabilidad:	
Nombre:	CrearExpReg (string clave, int pos)
Descripción:	Crea expresiones regulares teniendo en cuenta las etiquetas en formato HTML que lee del fichero de configuración y lo recibe por pase de parámetros a través de *clave* y con *pos* sabe la posición de dicha etiqueta en la página.
Nombre:	BuscarSeg
Descripción:	Método que machea el texto HTML en busca del segmento de nuestro interés, es decir, el que se encuentra entre las etiquetas señaladas, para extraerlo.
Nombre:	DevolverURL
Descripción:	Método que machea el texto HTML del segmento recuperado en BuscarSeg, hasta encontrar un hipervínculo, utilizando una expresión regular para extraer los enlaces a otras páginas de interés.
Nombre:	BorrarComent
Descripción:	

	Método que machea el texto utilizando una expresión regular para hallar los segmentos que están comentados y eliminarlos.

}

Tabla 3.8 Descripción de la clase AuxFunc

Nombre: Parser	
Tipo de clase: controladora	
Atributo	**Tipo**
PoliFunc	ArrayList
Metodos	AuxFunc
News	ArrayList
Inicial	Integer
Raiz	String
Nodo	TNodo
Info	TInfo
Para cada responsabilidad:	
Nombre:	Cargar_Pag (string dir)
Descripción:	Descarga las páginas correspondientes utilizando las clases que brinda .Net tras obtener los URLs de Internet mediante del método DevolverURL. Este URL lo recibe a través del parámetro *dir.*
Nombre:	Leer_Conf
Descripción:	Método encargado de leer del fichero de configuración y en dependencia de lo que va leyendo irá haciendo las llamadas necesarias para descargar la página e ir obteniendo las características de las noticias para luego asignarlas a los atributos que las recopilarán en la base de datos, de forma escalonada primeramente obtendrá el URL del sitio, tras cargar el home page, accederá a las páginas que nos interesan de dicho sitio, luego al segmento de esas páginas en las que aparezcan los hipervínculos con las noticias y por último a la noticia en sí.

Tabla 3.3 Descripción de la clase Parser

Nombre: TInfo	
Tipo de clase: entidad	
Atributo	**Tipo**
Clasif	String
Categ	String
NewsTex	String
raizNot	String
Para cada responsabilidad:	
Nombre:	-
Descripción:	-

Tabla 3.4 Descripción de la clase TInfo

Nombre: TNodo	
Tipo de clase: entidad	
Atributo	**Tipo**
raizEnl	String
Tipo	String
Textos	String
Para cada responsabilidad:	
Nombre:	-
Descripción:	

Tabla 3.5 Descripción de la clase TNodo

3.2.4 Diseño de la base de datos

Diagrama Entidad Relación de la base de datos.

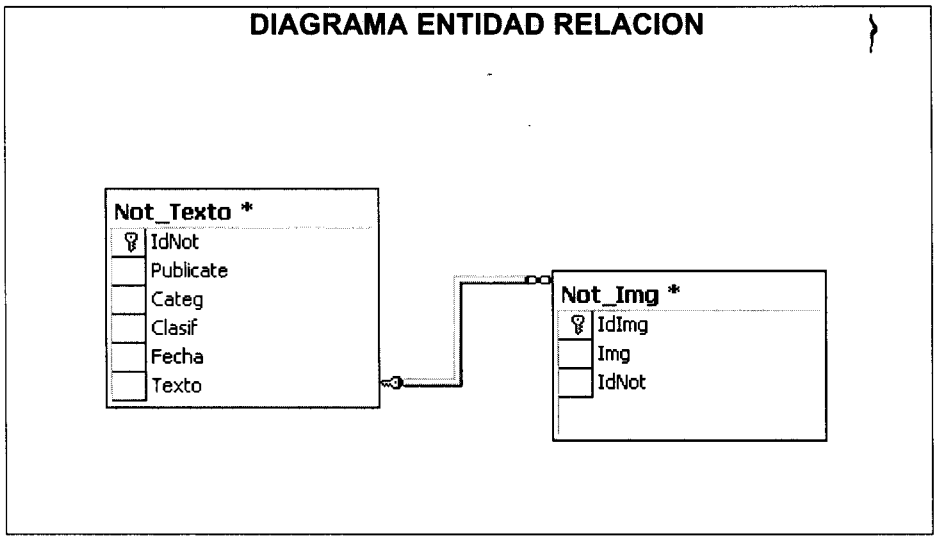


Fig. 3.8 Diagrama de Entidad – Relación

3.2.5 Descripción de las tablas.

Nombre: Not_Texto		
Descripción: Tabla en la quedarán almacenadas las noticias conjuntamente con algunas de las características esenciales que debemos tener en cuenta para su posterior publicación.		
Atributo	**Tipo**	**Descripción**
IdNot	int	Identificador de la noticia. Llave de la tabla. Atributo de gran importancia pues a través de él tenemos acceso a la noticia.
Publicate	bit	Atributo binario que hace referencia a la

		publicación o no de la noticia, 0 si no ha sido publicada y es el valor por defecto, 1 si ya fue publicada. Será utilizado cuando se trabaje con las publicaciones de las noticias.
Clasif	nvarchar	Atributo utilizado para señalar la procedencia de la fuente noticiosa: la clasificación de la noticia será nacional si es extraída del periódico Granma, y por otra parte, será internacional si se obtiene de CNN.
Categ	nvarchar	Atributo a través del cual se le da una categoría a la noticia atendiendo el tema al cual hace referencia, cultura, deporte, tecnología, salud, américa, etc.
Fecha	datetime	Atributo utilizado para almacenar la fecha en que fue descargada la noticia.
Texto	text	Atributo utilizado para almacenar el texto de la noticia.

Tabla 3.6 Tabla Not_Texto

Nombre: Not_Img		
Descripción: Tabla en la quedarán almacenadas las imágenes por cada una de las noticias.		
Atributo	**Tipo**	**Descripción**
IdNot	Int	Identificador de la noticia. Llave de la tabla Not_Texto. Atributo de gran importancia pues a través de él, sabemos a que noticia pertenece cada imagen. En la tabla puede estar repetido el valor de este atributo, pues se mantendrá para cada una de las imágenes que posea el artículo.
IdImg	Int	Identificador de la imagen. Llave de la tabla Not_Img. Atributo de gran importancia, a través de él tenemos acceso la noticia.
Img	image	Atributo que contiene la imagen.

Tabla 3.7 Tabla Not_Img

3.4 Conclusiones

En este capítulo se han detallado las ideas y los procedimientos utilizados para armar el esqueleto de nuestra aplicación. Con el objetivo de lograr, tras el ensamblaje de las piezas, el producto que deseamos.

Conclusiones

Nuestro sistema constituye un sencillo motor de búsqueda, con una funcionalidad bastante restringida, pues en su radio de acción solo maneja dos sitios de Internet: la versión digital del diario Granma nacional y CNN en español; ello trae consigo que no tengamos variedad de contenidos, pues solo brindamos noticias. Aun así, se conseguirá dar solución a un problema que presenta la Intranet de la Universidad de las Ciencias Informáticas, pues al brindar noticias de forma automática y única, el sistema se convierte en una solución muy eficiente para aquellos usuarios que deseen estar informados, y además quedarán almacenadas un tiempo prudencial por si se desea hacer referencia a algún suceso ocurrido en días anteriores.

Se han cumplido los requisitos y objetivos planteados, aportando un sistema que se apoya en un fichero de configuración, al cual referimos gran importancia, ya que almacena los datos de la ubicación exacta de los artículos que serán descargados de los sitios a visitar. De esta forma las búsquedas se realizarán de manera más precisa, y así no se desaprovecha el tiempo y se brinda la información concreta; siendo, la idea de contar con este fichero, que no es inherente al software, el principal aporte al mundo de los motores de búsqueda, pues actualizándolo, se evita realizar cambios en el código cuando existan variaciones en los sitios a consultar.

El uso de este sistema, nos asegura centralizar y distribuir la información noticiosa de la Universidad. Asegurando que los estudiantes, profesores, trabajadores, en fin, todo el personal que tenga acceso a la intranet, pueda contar con las herramientas necesarias para ser un ciudadano con conocimientos reales del acaecer mundial.

Conocemos la importancia que se le atribuye al hecho de estar informados de todo el acontecer, tanto nacional como internacional: "No podemos quejarnos. Nos ha tocado el privilegio de vivir lo que me atrevo a calificar como la más extraordinaria y decisiva época que ha conocido hasta hoy la especie humana.

(...) Por primera vez en la historia humana, nuestra especie corre un riesgo real de extinción. La amenazan no solo la destrucción de su medio natural de vida, sino también graves riesgos políticos, armas cada vez más sofisticadas de destrucción y exterminio masivo y doctrinas extremistas que podrían apoyarse en mortales y aniquiladoras fuerzas." [13]

Recomendaciones

Tras el desarrollo del sistema, y con el producto que hemos obtenido como resultado, nos quedan detalles que consideramos facilitarían el uso de esta aplicación y la harían más rica en contenidos. Estos detalles constituyen nuestras recomendaciones:

- ✓ Realizar un estudio más detallado de otras fuentes noticiosas en Internet en aras de encontrar patrones de búsqueda generalizados.
- ✓ Proveer al Sistema de una herramienta para facilitar la realización del fichero de configuración.
- ✓ Desarrollar de un Servicio Web XML que proporcione las funciones necesarias para utilizar el sistema e implementar aplicaciones que lo utilicen.
- ✓ Desarrollar de una aplicación administrativa que permita a un supervisor o editor aprobar o descartar las noticias para su posterior publicación.
- ✓ Proveer al Servicio Windows de un mecanismo para configurar datos como:
  - ❖ Servidor Proxy
  - ❖ Usuario
  - ❖ Contraseña
  - ❖ Puerto
  - ❖ Servidor de Base de Datos

Referencias Bibliográficas

- [01]. Lic. Ania Torres Pombertl. El uso de los buscadores en Internet. Disponible en URL: http://www.bvs.sld.cu/revistas/aci/vol11_3_03 (11/03/2003)

- [02]. Buscadores. (02/03/2004) Disponible en URL: http://www.besafeonline.org/spanish/buscado.htm (2002)

- [03]. Buscadores automáticos internacionales e hispanos. (11/06/2004) Disponible en URL: http://www.unav.es/fcom/guia/recursos/fr_3recursos_datos.htm (2003)

- [04]. Antonio Caravantes. Google News: las mejores noticias, automáticas, y gratis. (11/06/2004). Disponible en URL: http://www.caravantes.com/arti02/goglenew.htm (07-10-2002).

- [05]. José Antonio Robles Ordóñez. WWW ROBOTS (02/03/2004) Disponible en URL: http://polaris.lcc.uma.es/~eat/services/robots.html (1997)

- [06]. Robots de Motores de Búsqueda. Como Trabajan y Que Hacen. (02/03/2004). Disponible en URL: http://www.latin-marketing.com/boletin_posicionamiento/boletin-15-05-2003.htm (15/05/2003)

- [07]. Una introducción a los Robots de la World Wide Web. (02/03/2004) Disponible en URL: http://www.ldc.usb.ve/~redes/Temas/Tema.55/parte1.html (2000)

- [08]. M. en C. Armando F. Ibarra. Rational Unified Process. (20/04/2004)

- [09]. Discurso pronunciado por el Presidente de la República de Cuba Fidel Castro Ruz, en la sesión extraordinaria de la Asamblea Nacional del Poder Popular. Palacio de las Convenciones, 26 de junio del 2002.

- [10]. Julio Ariel Hurtado Alegría. "El lenguaje unificado del modelado". Universidad del Cauca. (2000)

- [11]. Javier Sánchez Pérez. "El lenguaje unificado del modelado. Metodologías de desarrollo software". (2000)

- [12]. Larman, Craig. "UML y patrones. Introducción al análisis y diseño orientado a Objetos". México. 1999

- [13]. Discurso pronunciado por el Presidente de la República de Cuba Fidel Castro Ruz, en la clausura del V Encuentro sobre Globalización y Problemas del Desarrollo, en el Palacio de las Convenciones, La Habana, el 14 de febrero del 2003.

Bibliografía

- [ARA02]. Aramburu, M.J., Sanz, I., García, S., "Procesamiento de periódicos electrónicos para su almacenamiento con Oracle 8" Dpto. de Informática. Universidad JAume I, Castellón. (2002)

- [BLA03]. Blanco Cuaresma, Sergio. "Mono: La plataforma .NET libre". http://www.marblestation.com/publicaciones/paper-mono.pdf (11/11/2003)

- [CAS01]. Castells, Pablo, Saiz, Francisco. "La web semántica: tecnologías y aplicaciones". Escuela Politécnica Superior. Universidad Autónoma de Madrid, (2001)

- [DEU02]. de Ugarte, David. "El Libro del posicionamiento en buscadores" (2002).

- [FIL98]. Filiberto, Franco Luis. "Motores de búsqueda" (1999) http://www.monografias.com/trabajos/buscadores/buscadores.shtml#Motordebusqueda

- [GUE01]. Guerrero, Luis A. "Rational Unified Process". Universidad de Chile. Dpto. Ciencias de la Computación. (2001)

- [GON] Gonzalez S., José Antonio. "El lenguaje de Programación C#"

- [IBA00] M. en C. Ibarra, Armando F. "Rational Unified Process"

- [MAR]. Martin, Richard. "SQL SERVER 2000 Databases for .NET Enterprise Server"

- [RAM01]. Ramírez, Javier A. "Motores de Búsqueda en Internet". Universidad Nacional de Luján. Argentina. (2001)

- [ROB97]. Robles, José Antonio Ordóñez. "WWW ROBOTS" http://polaris.lcc.uma.es/~eat/services/robots.html (1997)

- [PRE97]. Pressman, Roger. "Ingeniería de software. Un enfoque práctico", McGraw-Hill. Cuarta Edición. (1997)

- [ROD02]. Rodriguez, Daniel "Tutorial de Expresiones Regulares" http://bulma.net/body.phtml?nIdNoticia=770 (2002)

- [TOR03]. Lic. Torres Pombertl, Ania. "El uso de los buscadores en Internet". www.bvs.sld.cu/revistas/aci/vol11_3_03 (2003)

- [VIL02]. Viloria Lanero, Alejandro. "Introducción a las Expresiones Regulares. Teoría de Autómatas y Lenguajes Formales". 2002

- "Buscadores". http://www.besafeonline.org/spanish/buscado.htm#pagetop (2002)

- "Cómo funcionan los buscadores" http://www.alojarte.com/modules/news/article.php?item_id=2 (2003)

- "Creación de un buscador Web" http://www.towercom.es/nsp00016.html (1999)

- "Cuál es la razón de ser de una Base de Datos? http://www.osmosislatina.com/aplicaciones/bases_de_datos.htm (2003)

- "Expresiones regulares". http://www.ciberdroide.com/misc/novato/curso/regexp.html

- "Información general de ADO.NET" http://www.monografias.com/trabajos14/informe-ado-net/informe-ado-net.shtml (2003)

- "Introducción a las expresiones regulares"

http://www.lpsz.org/articulos/introduccion_expresiones_regulares.html
(2003)

- Lenguaje Unificado de Modelado(UML)
http://buscador.lycos.es/searchFrame/searchframe.html?url=http%3A%2F%2Fcfrela.en.eresmas.com%2F&query=uml&SITE=es&cat=loc&qstr=family%3Doff%26pic_adv%3D1%26pic_family_link%3Doff%26query%3Duml%26cat%3Dloc%26x%3D30%26y%3D15

- "Los servicios en Microsoft Windows"
www.wininfo.com.ar/main.html
(2001)

- "Microsoft SQL Server 2000 Enterprise Edition"
http://www.microsoft.com/spain/servidores/sql/default.asp
(2003)

- Popkin SOftware and Systems. "Modelado de sistemas con UML"
http://buscador.lycos.es/searchFrame/searchframe.html?url=http%3A%2F%2Flucas.hispalinux.es%2FTutoriales%2Fdoc-modelado-sistemas-UML%2Fdoc-modelado-sistemas-uml.pdf&query=uml&SITE=es&cat=loc&qstr=family%3Doff%26pic_adv%3D1%26pic_family_link%3Doff%26query%3Duml%26cat%3Dloc%26x%3D30%26y%3D15

- "¿Qué es C# ?"
http://www.desarrolloweb.com/articulos/561.php
(2003)

- "Robots de Motores de Búsqueda. Cómo Trabajan y Qué Hacen".
http://www.latin-marketing.com/boletin_posicionamiento/boletin-15-05-2003.htm
(2003)

- "Servicios de Windows"
http://www.microsoft.com/spanish/msdn/articulos/default.asp
(2004)

- "Sitio de XML"
http://www.xml.com.ve/

- "SQL Server"
  http://www.monografias.com/trabajos14/sqlserver/sqlserver.shtml

- "TÉCNICAS DE LOS MOTORES DE BÚSQUEDA BASADOS EN CRAWLERS"
  http://www.latin-marketing.com/boletin_posicionamiento/boletin-15-01-2003.htm
  (2003)

- Tutorial de ADO.NET
  http://es.gotdotnet.com/quickstart/aspplus/doc/quickstart.aspx
  (2002)

- Tutorial XML
  http://www.dat.etsit.upm.es/~abarbero/curso/xml/xmltutorial.html
  (1999)

- Una introducción a los Robots de la World Wide Web.
  http://www.ldc.usb.ve/~redes/Temas/Tema.55/parte1.html
  (2000)

Anexos

Anexo1.

Fichero de configuración XML.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <!-- Fichero de Portada -->
- <conf>
  - <sites clasif="internacional">
    - <!-- CNN Mundo -->
    <url>http://www.cnnenespanol.com/</url>
      - <step>
        - <!-- Ubicación de enlaces a paginas -->
        <operation tag="TD" cant="15" />
          - <!-- Enlaces Especificos a paginas -->
          <tipo categ="mundo" />
          <operation tag="A" cant="1" />
      </step>
      - <step>
        - <!-- Ubicación de enlaces a noticias -->
        <operation tag="TD" cant="55" />
          - <!-- Enlaces Especificos a noticias -->
          <operation tag="A" cant="2" />
          <operation tag="A" cant="3" />
          <operation tag="A" cant="4" />
      </step>
      - <step>
        - <!-- Ubicación de la noticia -->
        <operation tag="TD" cant="42" />
```

```
            </step>
</sites>

<sites clasif="internacional">
  - <!-- CNN AMERICAS, DEPORTES, TECNOLOGIA -->
    <url>http://www.cnnenespanol.com/</url>
          - <step>
                - <!-- Ubicación de enlaces a paginas -->
                  <operation tag="TD" cant="15" />
                        - <!-- Enlaces Especificos a paginas -->
                          <tipo categ="america" />
                          <operation tag="A" cant="2" />
                          <tipo categ="deportes" />
                          <operation tag="A" cant="6" />
                          <tipo categ="tecnologia" />
                          <operation tag="A" cant="7" />
                          <tipo categ="salud" />
                          <operation tag="A" cant="8" />
                          <tipo categ="cultura" />
                          <operation tag="A" cant="9" />
            </step>
          - <step>
                - <!-- Ubicación de enlaces a noticias -->
                  <operation tag="TD" cant="45" />
                        - <!-- Enlaces Especificos a noticias -->
                          <operation tag="A" cant="2" />
                          <operation tag="A" cant="3" />
                          <operation tag="A" cant="4" />
            </step>
          - <step>
```

```
            <!-- Ubicación de la noticia -->
            <operation tag="TD" cant="42" />

        </step>

    </sites>

- <sites clasif="nacional">

    - <!-- Granma -->

    <url>http://www.granma.cubaweb.cu/2003/07/08/menuizq.htm</url>

        - <step>

            - <!-- Ubicación de enlaces a paginas -->

            <operation tag="TABLE" cant="2" />

                - <!-- Enlaces Especificos -->

                <tipo categ="nacional" />

                <operation tag="A" cant="2" />

                <tipo categ="cultura" />

                <operation tag="A" cant="4" />

                <tipo categ="deporte" />

                <operation tag="A" cant="5" />

                <tipo categ="tecnologia"/>

                <operation tag="A" cant="8"/>

        </step>

        - <step>

            - <!-- Ubicación de enlaces a noticias -->

            <operation tag="TABLE" cant="3" />

                - <!-- Enlaces Especificos -->

                <operation tag="A" cant="1" />

                <operation tag="A" cant="2" />

                <operation tag="A" cant="3" />

        </step>
```

```
        -<step>
            - <!-- Ubicación de la noticia  -->
              <operation tag="TABLE" cant="3" />
          </step>
    </sites>
    <ontime time="06:00" />
</conf>
```

Anexo 2

Expansión del caso de uso # 1

Caso de uso	
CU-1	Iniciar el Servicio de Windows
Propósito	Dar inicio a la búsqueda, descarga y almacenamiento de las noticias.
Actores Sistema Operativo (actor abstracto)	
Resumen: El Servicio se encuentra corriendo en background leyendo del fichero de configuración, a una hora especificada en este, el sistema operativo se encarga de iniciarlo y el módulo comienza la búsqueda, descarga y almacenamiento de la información.	
Referencias	RF- 5

Acción del actor	**Respuesta del sistema**
1- El Sistema inicia el Servicio	
	2- El módulo comienza a leer del fichero de configuración XML los datos necesarios para ir descargando las noticias de Internet y almacenándolas en la base de datos.
	3- Ver caso de uso # 2
	4- Ver caso de uso # 3
	5- Ver caso de uso # 4

Precondición
Debe existir el fichero de configuración a través del cual el sistema irá realizando cada una de las acciones.

Anexo 3

Expansión del caso de uso # 2

Caso de uso	
CU-2	Extraer Contenido
Propósito	Extraer el contenido de las páginas a analizar, especificado en el fichero de configuración.
Actores: Sistema Operativo (actor abstracto)	
Resumen: El Servicio, utilizando los datos que contiene el fichero de configuración y apoyándose en el trabajo con expresiones regulares descarga la información.	
Referencias	RF- 1, RF – 2

Acción del actor	**Respuesta del sistema**
	1- Lee del fichero de configuración
	2- Ubica el URL del sitio a visitar
	3- Carga la portada del sitio
	4- Limpiar el texto descargado
	5- Crear expresión regular para acceder a las páginas que deseamos de la portada.
	6- Ubica los URL de las distintas páginas a visitar en le sitio.
	7- Crear expresión regular para acceder los hipervínculos dentro de

	una página específica.
	8- Ubica los URL de las noticias a descargar en las páginas visitadas.
	9- Crear expresión regular para acceder al hipervínculo de una noticia en particular.
	10 - Ubica la noticia en sí.
	11- Clasifica la noticia.

Anexo 4

Expansión del caso de uso # 3

Caso de uso	
CU-3	Analizar Contenido
Propósito	Brinda una serie de funciones auxiliares, tanto para la descarga como para la preparación de la noticia para su posterior almacenamiento.
Actores: Sistema Operativo (actor abstracto)	
Resumen: Crea las expresiones regulares para realizar el macheo con las páginas y realizar la descarga de los artículos, además tiene métodos para devolver los URL y otros para alistar la información antes de ser recopilada.	
Referencias	RF- 1, RF – 3

Acción del actor	**Respuesta del sistema**
	1- Crea expresiones regulares
	2- Buscar segmento de las páginas en los que se encuentran los URL
	2- Devuelve URL
	3- Elimina comentarios y churres del texto descargado.

Anexo 5

Expansión del caso de uso # 4

Caso de uso	
CU-4	Almacenar Contenido
Propósito	Recopila la información procesada y la inserta en una base de datos, además utilizando expresiones regulares, detecta la presencia de imágenes en el texto y las recopila también.
Actores: Sistema Operativo (actor abstracto)	
Resumen: Mediante expresiones regulares el sistema detecta la presencia de imágenes en los textos, además inserta en la base de datos toda la información extraída.	
Referencias	RF- 4

Acción del actor	**Respuesta del sistema**
	1-Inicia un trabajo con las noticias
	2- Obtener características de las noticias
	3- Inserta los artículos en la Base de Datos
	4- Inicia trabajo con las imágenes
	5- Ubica URL de las imágenes utilizando expresiones regulares.
	6- Modifica el camino de las imágenes
	7- Inserta las imágenes en la Base de

	Datos

Glosario de términos y siglas

.NET Framework: El Framework de .Net es una herramienta sencilla y potente para distribuir el software permitiendo crear aplicaciones sólidas en forma de servicios que puedan ser suministrados remotamente y que puedan comunicarse y combinarse unos con otros de manera totalmente independiente de la plataforma, lenguaje de programación y modelo de componentes con los que hayan sido desarrollados, por lo que reducen extraordinariamente el desarrollo de aplicaciones. Aunque es posible escribir código para la plataforma .NET en muchos otros lenguajes como Visual Basic .NET, C++, J#, etc.; C# es el único que ha sido diseñado específicamente para ser utilizado en ella, por lo que programar usando C# es mucho más sencillo e intuitivo que hacerlo con cualquiera de los otros lenguajes ya que C# carece de elementos heredados innecesarios en .NET. Por esta razón, se suele decir que C# es el lenguaje nativo de .NET.

Actor: Es la persona que interactúa con el sistema o negocio.

ADO: ActiveX Data Objects. Es una tecnología ampliable y de fácil uso para agregar acceso a bases de datos a las páginas Web

Atributo: Es un valor lógico de un estado de un objeto.

Buscadores: Los buscadores son bases de datos creadas a partir de la información obtenida por programas que sistemáticamente recorren la Internet, localizando recursos de información y permitiendo su interrogación por palabra clave.

Common Language Runtime(CLR): Entorno de tiempo de ejecución que proporciona .NET Framework, que ejecuta el código y proporciona servicios que facilitan el proceso de desarrollo.

Hipervínculos: También denominado enlace, hiperenlaces o link, en inglés, se trata de las especificaciones que permiten saltar de un documento a otro dentro de un sitio web o dentro de toda la Red, sólo con pulsar sobre él. Puede ser un texto o una imagen.

Home Page : Se refiere a la pantalla principal de un Web site; es como la unión de la tapa de un libro y su tabla de contenidos. Pero tal como en los libros

donde los podemos abrir en cualquier página, no es necesario comenzar desde la home page. Se puede consultar cualquier parte de la Web site.

Internet: Conjunto global de redes de ordenadores conectados entre sí; el intercambio de información entre las diversas redes se realiza mediante protocolos TPC/IP que fueron desarrollados por ARPANET a finales de los sesenta. De manera básica se puede decir que Internet lo conforman cientos de computadoras que están comunicados entre sí y que son de dominio público. Estos equipos contienen las páginas html (Hypertext markup lenguaje) que son las responsables de lo que vemos en el browser o navegador.

Intranet: Redes tipo Internet pero que son de uso interno, por ejemplo, la red corporativa de una empresa que utilizara protocolo TCP/IP y servicios similares como www.

Multilenguaje: Que admite más de un lenguaje.

Motor de búsqueda: Un motor de búsqueda es un servicio o un conjunto de programas coordinados que se encargan de visitar cada uno de los sitios que integran el Web, indexa, organiza y a menudo califica y revisa sitios Web empleando los propios hipervínculos contenidos en las páginas, para luego presentar direcciones en Internet como resultado de las peticiones de búsqueda solicitadas por usuarios que usan estos servicios de localización de páginas. Esto ayuda a encontrar la aguja que anda buscando en el pajar de Internet.

Procesador: Un microprocesador es un circuito electrónico integrado que actúa como unidad central de proceso de un ordenador, proporcionando el control de las operaciones de cálculo.

Robot: Manipulador multifuncional automático controlado, reprogramable, utilizado para la ejecución de tareas variadas. Normalmente su uso es el de realizar una tarea de manera cíclica.

Sistema: Software que controla el ordenador. Se toma la tarea de cargar otros programas y ejecutarlos y provee acceso a los archivos.

Sitio Web: Es un conjunto de archivos electrónicos y páginas Web referentes a un tema en particular, que incluye una página inicial de bienvenida,

generalmente denominada home page, con un nombre de dominio y dirección en Internet específicos.

Software: Conjunto de programas, instrucciones y reglas informáticas para ejecutar ciertas tareas en una computadora. Es un término o palabra en inglés que significa aplicación.

UML: Lenguaje para construir modelos; no guía al desarrollador en la forma de realizar el análisis y diseño ni le indica cual proceso de desarrollo adoptar. (D)

URL: Uniform Resource Locators. Cada documento y recurso en internet tiene una dirección única, conocida como su localizador de recursos uniforme.

Web Crawler: Robot que recopila páginas web para el índice de los motores de búsqueda.

World Wide Web (WWW): World Wide Web, o simplemente Web, es el universo de información accesible a través de Internet, una fuente inagotable del conocimiento humano.

XML: Es un lenguaje de metamarcado que nos ofrece un formato para la descripción de datos estructurados.