Herramientas para Consulta y Modelado en la Web, una forma diferente del manejo de grandes volúmenes de información de los Web Sites en Internet*

Daladier Jabba Molinares** y José Márquez Díaz***

Resumen

La Web se ha convertido en uno de los principales medios para publicar información. Esto ha traído como consecuencia que grandes volúmenes de datos se tengan que manipular y desplegar de una forma ágil y a la vez agradable al usuario, por lo que el despliegue de éstos en sus respectivas páginas se convierte ahora en un problema para los administradores de sitios Web no sólo por el montaje sino también por su mantenimiento. Existen herramientas que permiten diseñar y en general modelar sitios Web de una forma más rápida y dinámica intereactuando con bases de datos, lo cual permite a los administradores realizar una labor más eficiente. El resultado de esta investigación tiene como fin mostrar algunas de estas herramientas que existen en la Web para consulta y modelado, como lo son Strudel y Araneus.
Palabras clave: http, web, Strudel, Araneus, servidores Web, repositorios de datos.

Abstract

The Web has been become in one of the main media to publish information, this has got as consequence that a high data volume have to be manipulated and displayed in agile manner and, at the same time, enjoyable to the user; for this reason, the display of data has become in a problem for Web Site Administrators because of the upload of the pages and their maintenance. There exist tools, which allow to design and to model, in general, a web site in a quickly and dynamic way, interacting with databases; this permit that the administrator can be more efficient in his/her tasks. The result of this investigation has as purpose to show some of these tools that exist in the web for requesting and modeling such as Strudel, and Araneus.
Key words: http, web, Strudel, Araneus, Web Server, Data Repository.

Fecha de recepción: 19 de julio de 2002

* Este artículo forma parte de los resultados de la investigación *Consulta y modelado en la Web,* del Grupo de Redes de Computadores, línea de investigación: Desarrollo de aplicaciones para la web. Universidad del Norte

** Ingeniero de Sistemas de la Universidad del Norte; Magíster en Ciencias Computacionales del convenio ITESM-CUTB. Docente tiempo completo del Departamento de Sistemas de la Universidad del Norte. *djabba@uninorte.edu.co*

*** Ingeniero de Sistemas de la Universidad del Norte; Magíster en Ciencias Computacionales del convenio ITESM-CUTB. Docente tiempo completo del Departamento de Sistemas de la Universidad del Norte. *jmarquez@uninorte.edu.co*

INTRODUCCIÓN

Internet ofrece un servicio de búsqueda de información avanzado a muchas computadoras. Este servicio, conocido como *World Wide Web* (www), enlaza y reúne la información almacenada en muchas computadoras. El creciente desarrollo de los sitios Web ha impuesto un desafío en el problema de las bases de datos. Esto ha generado un número significativo de propuestas de investigación en las áreas de bases de datos para la administración de sitios Web; otros trabajos importantes en este campo han investigado la extensión de las metodologías para el diseño de estos sitios y su interacción con las herramientas de desarrollo.

En el www existen numerosos sitios, cuyo manejo de contenido y estructura es un nuevo problema que no había sido tenido en cuenta para los que administran bases de datos. Las principales tareas de las personas que están encargadas de la construcción de sitios Web son: elección y acceso a los datos que se desplegarán en el sitio, donde se especifican los datos contenidos dentro de cada página y los enlaces entre páginas; y el diseño de la presentación visual de las páginas. Actualmente estas herramientas son muy independientes.

Las tareas importantes de la Web, tales como la actualización automática de un sitio, la reestructuración del sitio y las restricciones de integridad en un sitio determinado se realizan de una manera tediosa. Además de todo lo mencionado anteriormente, se presenta quizás uno de los más grandes problemas en la Web, como son los volúmenes de información que se manejan, los cuales son demasiado grandes y difíciles de controlar, por lo que se hace necesario involucrar las técnicas de bases de datos para poder afrontar este problema y poder facilitar el manejo de la información a los usuarios que quieran acceder a ella, ya que en general las técnicas de bases de datos tienen como objetivo facilitar el manejo y control de grandes volúmenes de datos, y estas técnicas se pueden adoptar como solución a las necesidades que se presentan en la Web. Si se ve a la Web como un gran grafo, es natural plantear consultas que vayan más allá del paradigma de recuperación de la información básica soportado por los motores de búsqueda. Es necesario tener en cuenta tanto la estructura interna de las páginas Web como la estructura externa de los enlaces que las interconectan.

A través de este estudio se hace una revisión bajo la óptica de la arquitectura, el modelo de datos y los lenguajes que se utilizan para llevar a cabo la solución a los problemas generados por el creciente cambio en los contenidos y estructuras de información de los sitios Web. Estas herramientas permitirán mantener actualizado el sitio con la información que se genere dinámicamente en una organización. La importancia de éstas radica en que el usuario que mantiene un sitio no tendrá que preocuparse por los detalles de los contenidos, ni la estructura del sitio, ya que la información se almacenará en repositorios de datos, los cuales deben estar actualizados para que la información que se despliegue también esté al día.

1. MANEJO DE INFORMACIÓN EN LA WEB

1.1. ¿Qué es Web?

La Web fue diseñada sobre Internet y utiliza TCP/IP para transferir información de un lugar a otro. Los usuarios (llamados clientes) utilizan programas exploradores, como el Internet Explorer o Navegator de Netscape, para conectarse con servidores Web, FTP y nodos Gopher, o para enviar y recibir correo electrónico. Una vez realizada la conexión con Internet, se puede utilizar un explorador y saltar de nodo Web en nodo Web.

1.2. Administración de la información en la Web

Existen tres clases de tareas que se relacionan con la WWW: modelamiento y consulta en la Web, extracción de información e integración, construcción y reestructuración de sitios, las cuales serán descritas en los siguientes aspectos [5]:

1.2.1. **Modelamiento y consulta en la Web**

Se mira a la Web como un grafo dirigido cuyos nodos son páginas Web y cuyas aristas son los enlaces entre las páginas. Las consultas están basadas en el contenido de las páginas deseadas y en la estructura de enlaces que conectan las páginas.

La manera más simple de resolver esta tarea la realizan los motores de búsqueda en la Web para la localización de páginas basada en las palabras que ellos contienen.

1.2.2. **Extracción de información e integración**

Ciertos sitios Web pueden ser vistos como contenedores de estructuras de datos. Dado el alcance en el número de tales sitios, se pueden considerar dos tareas. La primera es extraer una representación estructurada de los datos de las páginas HTML que las contienen; segunda tarea que se debe ejecutar dentro de estos sitios es el planteamiento de consultas que requieren de la integración de datos.

1.2.3. **Construcción y reestructuración de sitios Web**

La tecnología de bases de datos también se pueden aplicar en la construcción, reestructuración y manejo de sitios Web. Al contrario de los dos casos anteriores, que aplican sobre sitios Web existentes, éste consiste en la creación de sitios Web, ya sea a partir de datos puros (almacenados en bases de datos o en archivos estructurados) o reestructurando los sitios Web existentes.

2. REPRESENTACIÓN DE DATOS PARA LA WEB

Los sistemas de construcción para solución de algunas de las tareas planteadas inicialmente requieren que se elija un método para el modelamiento, entre los cuales están el modelo de grafo de datos y el modelo de datos semiestructurados, los cuales serán descritos a continuación [5].

2.1. **Modelo de grafo de datos**

Varias de las aplicaciones requieren modelar el conjunto de páginas Web y los enlaces entre ellas. Estas páginas pueden estar en varios sitios o dentro de un sitio simple. En este modelo, los nodos representan páginas Web (o componentes internos de las páginas Web), y los arcos representan los enlaces entre las páginas Web, las etiquetas en los arcos pueden ser vistas como atributos.

2.2. **Modelo de datos semiestructurados**

El segundo aspecto de modelamiento de datos para aplicaciones Web se basa en que la estructura de los datos en muchos casos es irregular; cuando se modela la estructura de un sitio Web no se tiene un esquema fijo que se dé por adelantado. El modelamiento de datos que se toma de distintas fuentes implica que la representación de algunos atributos pueda diferir de fuente a fuente; de allí que se plantee la necesidad de considerar modelos de datos semiestructurados[5].

3. MODELOS DE LENGUAJES DE CONSULTA EN SITIOS WEB

Para construir un sitio Web es necesario cubrir las siguientes tareas:

- Seleccionar y manejar los datos que van a estar disponibles en el sitio Web
- Organizar la información en páginas individuales o en grafos de páginas enlazadas
- Diseñar la presentación visual de las páginas
- Diseñar la base de datos necesaria para el manejo de la información que va a estar al alcance en la Web.

3.1. CLASIFICACIÓN DE SITIOS WEB

En Internet existen cuatro formas de clasificar los sitios de la Web, de acuerdo con su complejidad en términos de datos y aplicaciones [8]:

3.1.1. **Sitios de presencia en la Web**

Los sitios de presencia en la Web tienen baja complejidad tanto en términos de datos

como de aplicaciones; éstos contienen un pequeño número de páginas, y se utilizan para propósitos de mercadeo. Estos sitios generalmente son desarrollados manualmente con la ayuda de editores HTML (Lenguaje de Marcas Hipertextuales) y de programas muy simples para su manejo.

3.1.2. **Sitios orientados al servicio**

Los sitios orientados al servicio se dedican muy especialmente a algún tipo de servicio. Entre éstos están los motores de búsqueda y los servicios de correo electrónico gratuitos. En estos sitios la estructura de los datos y del hipertexto son muy simples, y su complejidad se fundamenta en las aplicaciones que garantizan el servicio.

3.1.3. **Sitios de datos intensivos**

Los sitios de datos intensivos publican gran cantidad de información, y por lo tanto tienen una estructura muy compleja de hipertexto, y ofrecen poco o ningún servicio. Entre estos sitios están los académicos, los cuales publican cursos e investigaciones. El enfoque principal consiste en organizar los datos en un formato de hipertexto, y en el mantenimiento de estos datos y su formato de hipertexto.

3.1.4. **Sistemas de información basados en la Web (WBIS)**

Los sitios orientados a sistemas de información basados en la Web (WBIS) presentan como mejor ejemplo los sistemas de información reales en la Web y ofrecen acceso a datos complejos y, al mismo tiempo, también proporcionan servicios interactivos sofisticados. Los sitios que comercializan electrónicamente son los que constituyen esta categoría, tales como los sistemas de información que se basan en plataformas Internet.

3.2. DESARROLLO EN LA WEB [6]

A partir del ciclo de vida de un proceso tradicional se pueden deducir las etapas del ciclo de vida desarrollado para la Web. Estas etapas son las siguientes:

- *Análisis de Requerimientos:* Se identifican los requerimientos sin perder de vista que serán aplicaciones con posibilidades de acceso universal. Se debe definir la naturaleza de la información que se va a publicar.

- *Conceptualización:* La aplicación se representa a través de un conjunto de modelos abstractos. La preocupación más importante es cómo aparecerán los objetos y las relaciones con los usuarios antes que su representación dentro del *software.*

- *Prototipo y Validación:* El prototipo debe ser construido con anterioridad al diseño y con una arquitectura simplificada; contiene un conjunto de páginas muestras que emulan la apariencia de lo que será la aplicación futura.

  - Diseño: Aquí se diferencian las siguientes vistas:
  - La vista estructural: La cual se hace corresponder con el esquema de almacenamiento
  - La vista navegacional: Conjunto de primitivas de acceso sobre el esquema de almacenamiento
  - La vista presentacional: Conjunto de especificaciones visuales (estilos) de contenido independiente

- *Implementación:* La aplicación es construida con nuevos contenidos preparados por expertos en el tema y/o con datos existentes de otros sistemas. Las páginas se construyen colocando los contenidos y los comandos de navegación en el estilo apropiado de presentación.

- *Evolución y Mantenimiento:* Después de su liberación, los cambios en las necesidades pueden requerir la revisión de la estructura, la navegación, la presentación o los contenidos. Los cambios son aplicados en una etapa muy superior y luego se propagan hacia abajo hasta la implementación.

4. HERRAMIENTAS PARA CONSULTA Y MODELADO EN LA WEB

Muchos sitios Web incluyen piezas sustanciales y significativas de información, de tal forma que son a menudo difíciles de construir, diseñar, correlacionar y mantener. Además debido a la popularidad de la *World Wide Web* (WWW), la necesidad de organizar grandes cantidades de datos en formas hipertextuales se ha incrementado. Para apoyar el proceso de diseño de estos grandes sitios Web y de gran intensidad de datos, se han propuesto recientemente varias metodologías en el contexto de diseño hipermedial. Esas metodologías representan un primer paso hacia la solución del problema, además proveen un modelo de datos específico y pasos metodológicos que asisten en el proceso de diseño de una aplicación hipermedial.

Sin embargo, el proceso de diseño también debe apoyar la actividad de mantenimiento de los sitios. En algunos servidores de índice Web, tales como Altavista, se puede ver un gran número de sitios que luego de ser establecidos son completamente abandonados, y presentan numerosos errores e inconsistencias debido al deficiente mantenimiento que se les presta, además su información es a menudo obsoleta.

Como consecuencia de este crecimiento se han creado nuevos tipos de herramientas como Strudel y ARANEUS para el manejo de problemas relacionados con la

construcción y el mantenimiento de sitios Web. Estos nuevos tipos de herramientas se fundamentan en la importancia que tienen los conceptos de bases de datos para los problemas de manejo y búsqueda de la información, su almacenamiento y mantenimiento.

4.1. STRUDEL

Strudel es un sistema que fue desarrollado conjuntamente por Mary Fernández (Laboratorios AT&T), Alon Levy (Universidad de Washington), Daniela Florescu (INRIA, Francia), Jaewoo Kang (Serera Inc.) y Dan Suciu (Laboratorio AT&T), con el fin de solucionar los problemas relacionados con la construcción y mantenimiento de los sitios Web. Fue el primer sistema en aplicar conceptos de los manejadores de bases de datos para la solución de estos problemas [3].

La idea principal del sistema Strudel [4] es la declaración de la estructura y el contenido del sitio Web en un lenguaje de consulta de alto nivel, StruQL *(Site Transformation Und Query Language).* Como resultado, se facilita la reestructuración y modificación de los sitios Web, así como también la creación de sus múltiples versiones a partir de los mismos datos básicos. La declaración de representaciones básicas de los sitios Web proporciona una plataforma para especificar y codificar la integridad sobre éstos *(constraints),* al igual que para el diseño de bodegas de datos para su soporte. La principal motivación para el desarrollo de Strudel es que la creación y mantenimiento de los sitios Web, con la actual tecnología, resulta tediosa, ya que éstos a menudo se derivan de múltiples fuentes de datos y tienen estructuras complejas. De allí que sus constructores necesiten de herramientas para la administración de sus contenidos y estructuras. El sistema Strudel aplica los conceptos de los sistemas de administración de bases de datos a los procesos de construcción de sitios Web. Con este sistema se realiza la separación entre la administración de los datos, la creación y administración de la estructura y la representación gráfica de las páginas del sitio.

Strudel se basa en un modelo de datos semiestructurados de grafos dirigidos etiquetados, los cuales se caracterizan por tener poca fortaleza, una estructura irregular y rápidamente evolucionan o pierden el esquema. Este modelo de datos fue adoptado por Strudel, ya que los sitios Web son grafos con estructura irregular y esquemas no tradicionales. Además, los datos semiestructurados permiten su integración desde múltiples fuentes no tradicionales. Con esta herramienta se han desarrollado, entre otros, los sitios Web de CNN y la compañía de automóviles Toyota. En la figura 1 podemos ver una muestra de la arquitectura de Strudel.

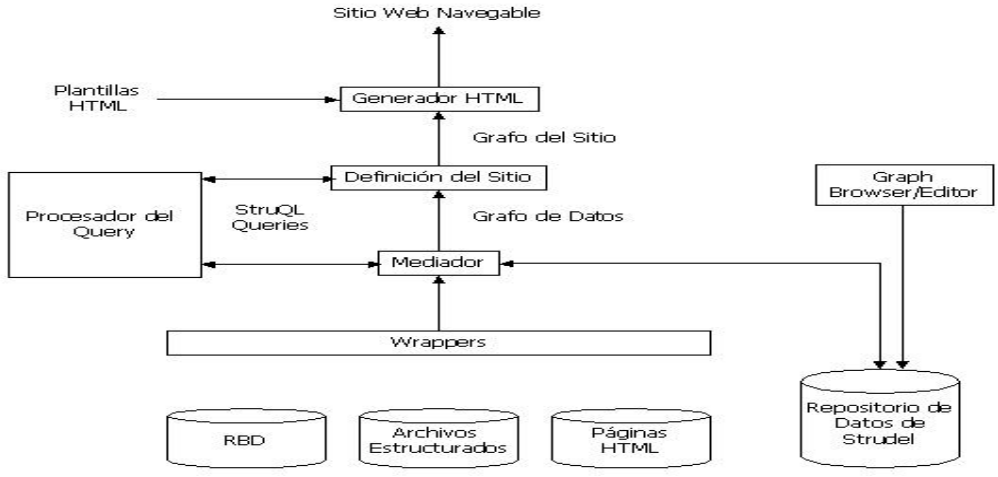

Figura 1. Arquitectura de Strudel

4.1.1. **Arquitectura**

A continuación se presenta una descripción (de abajo hacia arriba) de la función que cumple cada proceso dentro de la arquitectura Strudel y los tipos de fuentes (internas y externas) de donde es tomada la información que se mostrará en el sitio [3].

1. ***Fuentes de datos externas.*** Strudel utiliza como fuentes externas de suministro de datos para la construcción del sitio Web las bases de datos relacionales, las bases de datos orientadas a objetos, archivos estructurados (documentos de Word, Excel, texto o incluso algunos archivos HTML) y sitios Web existentes, que son traducidos a un modelo de grafo de datos por un *wrapper.*

2. ***Repositorio de datos para datos semiestructurados.*** Strudel cuenta con un repositorio propio de datos en el cual almacena los *grafos de datos* y *grafos del sitio* de un sitio Web. Estos datos también pueden ser obtenidos a partir de *wrappers* que transforman datos de fuentes externas en el formato interno de Strudel.

3. ***Wrappers.*** Son los encargados de traducir o convertir los datos que se encuentran en fuentes externas del modelo de grafos de Strudel. El proceso consiste en convertir los datos que se desea presentar en el sitio en un formato propio de Strudel, un grafo de datos, que es la representación de un modelo de datos semiestructurado. Esta representación se realiza a través de un grafo dirigido etiquetado.

4. ***Mediador.*** Una de las bondades de Strudel es su habilidad para construir sitios Web que se sirven de datos de múltiples fuentes. Esta funcionalidad está soportada en la herramienta Mediator, que proporciona una vista uniforme de todos los datos fundamentales, independiente de donde son almacenados. El diseño de Mediator tuvo en cuenta las siguientes consideraciones, que generalmente corresponden a las aplicaciones de integración de datos. La primera consideración es si ejecutar la integración de los datos por la aplicación de bodegas de datos de fuentes externas o acceder a las fuentes externas bajo demanda en tiempo de consulta. La segunda consideración es cómo especificar las relaciones entre los atributos y las colecciones en el esquema mediado y en las fuentes de datos.

5. ***Procesador de consultas.*** Strudel posee un nuevo lenguaje, StruQL, para las consultas y reestructuración de datos semiestructurados. Ya que los grafos de datos y de sitios son representados bajo el mismo modelo de datos (grafos etiquetados dirigidos), las consultas de StruQL pueden aplicarse a cualquier grafo, ya sea producido por un *wrapper*, una consulta de mediación o una consulta de definición de sitio.

6. ***Generador de HTML.*** Para producir la presentación gráfica de cada página en el sitio Web, Strudel asocia una plantilla HTML con cada nodo en el grafo del sitio. Las plantillas HTML pueden ser asociadas con colecciones de objetos en el grafo o con objetos individuales. Dado un objeto y su plantilla HTML, el generador interpreta la plantilla HTML, reemplazando las expresiones de la plantilla por los valores HTML de los atributos del objeto. Las páginas resultantes son el sitio Web por explorar.

7. ***Examinador/Editor de repositorio de datos.*** El examinador/editor del repositorio permite a un usuario crear, actualizar y ver grafos, y pueden ser ser utilizados tanto para grafos de datos como de sitios. Los usuarios pueden especificar selecciones simples en un objeto del grafo y ver los resultados. El editor también proporciona una interface amigable para la creación de nuevos grafos.

4.2. ARANEUS

El sistema ARANEUS fue desarrollado por Paolo Atzeni, Paolo Merialdo (Dipartamento di Informatica e Automaziane, Universidad de Roma Tre) y Giansalvatore Mecca (Universidad della Basilicata, DIFA). ARANEUS posee un proceso completo y sistemático de diseño para organizar y mantener grandes cantidades de datos en un hipertexto Web [1]. Es un sistema para manejar bases Web que consiste en una colección de datos de naturaleza heterogénea y muy sofisticada, es decir, datos altamente estructurados, tales como los almacenados en sistemas de bases de datos relacionales y orientados a objetos y datos semiestructurados con el estilo Web que incorporan bases de datos y sitios Web [7]. El Sistema Manejador de Bases Web

ARANEUS (WBMS) incorpora herramientas de consulta e integración de datos estructurados y semiestructurados que permiten el acoplamiento de la extracción de datos con la generación del hipertexto. Además, es parecido al DBMS en el manejo de tablas y el protocolo utilizado para comunicarse con la base de datos es el estándar SQL de JDBC. También introduce un número de herramientas y técnicas para manejo de bases Web, es implementado en Java y ejecutado en una plataforma Java compatible. Actualmente los sitios Web de las universidades de Roma y de la Basilicata, entre otros, han sido desarrollados bajo esta herramienta. En la figura 2 podemos ver un esquema de la arquitectura de ARANEUS.

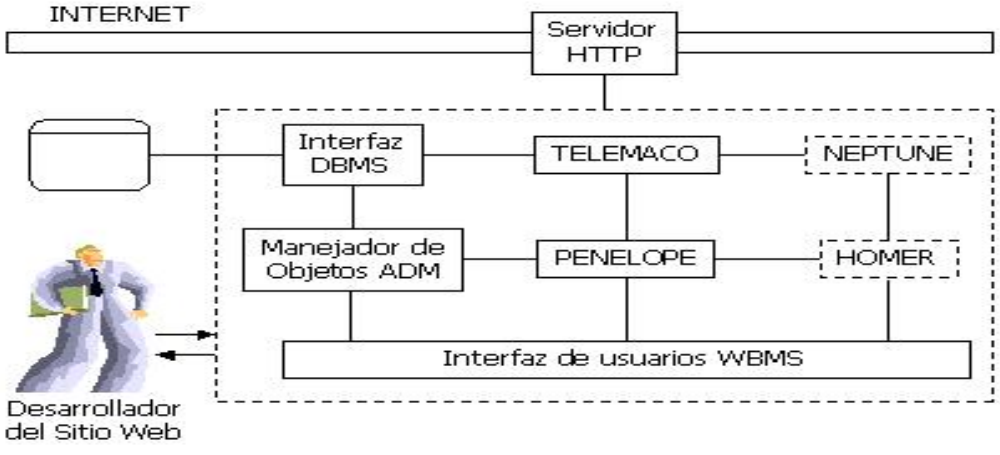

Figura 2. Arquitectura de ARANEUS

4.2.1. **Arquitectura**

Este sistema se caracteriza por poseer los siguientes componentes para el diseño, creación, implementación y mantenimiento de los sitios Web:

1. ***ADM (Modelo de Datos ARANEUS).*** Proporciona una descripción de un sitio Web y abstrae las características lógicas de las páginas Web. También permite dar una descripción lógica de la estructura, lo cual trae como consecuencia: primero, aislar cuidadosamente las características lógicas de las físicas, así se introduce una forma de hipertexto con independencia de datos, al igual que el utilizado en las bases de datos relacionales; segundo, el esquema lógico de hipertexto permite obtener una organización completa de las páginas en el sitio para evaluar la efectividad y eficiencia de la estructura escogida y posiblemente reestructurarla. Además abarca la descripción de los datos a distintos niveles como sigue:

- Al definir las técnicas de la consulta para la Web, se considera a la gran Web como una enorme base de datos.
- Considerar la Web como una colección de sitios, donde cada sitio es como una base de datos independiente; esto se justifica en el hecho de que los sitios Web son a menudo totalmente específicos en sus datos y servicios.
- Puede considerarse cada página HTML como una fuente de datos independiente.

El ADM representa un componente muy importante en el sistema. Cada página es vista como un objeto URL (identificador) anidado y un conjunto de atributos; éstos pueden ser simples, como texto, imágenes o sonidos; o complejos, como las listas de tuplas, posiblemente anidadas. Las páginas similares son agrupadas dentro de esquemas de páginas, los cuales toman la noción de esquema relacional o clases en bases de datos. Los esquemas de páginas son conectados utilizando enlaces, usados para describir las navegaciones en el sitio. Un sitio Web es una colección de esquemas de páginas conectados por enlaces.

2. ***Interfaz DBMS.*** Se basa en un lenguaje de consulta de SQL (Lenguaje Estructurado de Consultas) en el JDBC-ODBC (Conectividad Abierta de Base de Datos). Se utiliza para almacenar los objetos de la base de datos externa, descomponiéndola en tablas planas. Cuando se generan páginas HTML, se utiliza la interfaz para distribuir consultas en la base de datos y construir vistas de hipertexto. El aplicativo que genera sitios Web complejos utiliza la interfaz para emitir consultas a la base de datos y construir vistas de hipertexto que son luego traducidos en HTML. La principal ventaja de este esquema es que la página siempre refleja el estado más reciente de la base de datos [7].

3. ***Penélope.*** El módulo que apoya el proceso de definir y mantener los nuevos sitios se denomina PENÉLOPE [1], además alivia la carga del manejo de archivos HTML ante la presencia de actualizaciones y reorganizaciones. De esta forma, el diseñador puede concentrarse en la organización de hipertexto y utilizar esta herramienta para garantizar que el sitio esté consistentemente actualizado [7]. La fase de creación de sitios utilizando PENÉLOPE toma como entrada la descripción ADM del hipertexto objetivo, además del estilo de la página, generado por TELÉMACO. Este incorpora el álgebra de objetos ADM, un álgebra relacional anidada con el URL, que puede ser utilizada para generar vistas ADM y además páginas HTML sobre las tablas de la base de datos. La estructura del sitio Web objetivo y la correspondencia con la fuente de base de datos son descritas en el sistema por medio del Lenguaje de Definición PENÉLOPE (PDL), que es un lenguaje de declaración; adicionalmente, el Lenguaje de Manipulación PENÉLOPE (PML) permite crear páginas específicas, como también el sitio total. Cuando se genera un nuevo sitio, primero se diseña el esquema ADM. Este proceso es apoyado por un diseño de metodología específico. Luego la estructura ADM establece una correspondencia con la base de

datos utilizando PDL, el cual provee un comando para describir la estructura de un esquema de página en términos de las tablas de la base de datos; los esquemas de páginas pueden ser anidados arbitrariamente y enlazados utilizando el álgebra de objetos ADM. Luego se genera el sitio utilizando PML. La nueva página puede incluir enlaces a páginas existentes.

- *Lenguaje de Definición PENÉLOPE (PDL).* Una definición PDL específica cómo las páginas establecen correspondencias entre el sitio y la base de datos. Esto es realizado por un comando de definición en el sitio, seguido por una colección de definiciones de esquemas de páginas, uno para cada esquema de páginas en el sitio.

  Cuando el sistema PENÉLOPE ejecuta un comando PDL, primero se ejecuta una unión natural de las tablas especificadas en la cláusula USING. Luego todos los atributos URL necesarios se adicionan a la relación resultante. El valor del URL es generado por el operador URL desde de su parámetro actual (una cadena constante o un valor de la base de datos). Después se deshace de todos los atributos que no son útiles para la generación de instancias de páginas, es decir, aquellos atributos de la base de datos que no son asociados con atributos de página en el esquema de página. Finalmente, la relación resultante es transformada en un conjunto de tuplas anidadas: las operaciones anidadas son ejecutadas iniciando desde el interior de los atributos ADM del tipo LIST-OF. Cada tupla anidada de la relación resultante corresponde a una página en el sitio.

- *Lenguaje de Manipulación PENÉLOPE (PML).* La creación de páginas, definida en el código fuente de PDL, se ejecuta por medio de instrucciones del *lenguaje de manipulación PENÉLOPE* (PML). El algoritmo de mantenimiento de páginas toma como entrada una actualización de la base de datos, y retorna un conjunto mínimo de instrucciones necesarias para la correspondiente actualización de las páginas. En esencia, si se requiere una actualización a la base de datos del sistema, esto genera automáticamente una transacción mezclada, en la cual SQL actualiza las tablas de la base de datos y PML actualiza las páginas que son combinadas para garantizar la consistencia entre los dos. Luego automáticamente la transacción se ejecuta contra la base de datos y el sitio Web.

4. ***TELÉMACO.*** Una de las tareas más difíciles y subestimadas en desarrollos de sitios Web consiste en manejar la presentación gráfica de las páginas. Existe una lista de tres requerimientos fundamentales en este campo:

   - Primero, es muy útil tener prototipos de herramientas rápidas que producen algunas presentaciones aproximadas para todas las páginas en el sitio; esto permite concentrarse en otros aspectos del diseño de sitios.

- Luego, se deben tener herramientas flexibles para mejorar los detalles de la presentación y obtener como resultado una apariencia final.
- Finalmente, la presentación es desarrollada por código; es más conveniente trabajar en *ejemplos de páginas* HTML, que pueden ser desplegadas utilizando un navegador estándar para conseguir una reutilización inmediata, y luego permitir al sistema deducir desde los ejemplos el código necesario.

Estas ideas han inspirado a TELÉMACO, una herramienta para el diseño y desarrollo de la presentación. Cuando se construye un esquema de página con PENÉLOPE, los valores de los atributos son formateados según las instrucciones especificadas a través de TELÉMACO. Una noción fundamental es el *atributo estilo*, el cual especifica cómo los valores de un atributo dado pueden ser formateados en una página. Para poder producir un formato sofisticado, un atributo estilo se deduce de dos piezas arbitrarias de código, llamados *formato prefijo de cadena* y *formato sufijo de cadena*, entre los cuales los valores de los atributos son encapsulados cuando se generan páginas (Para algunos tipos de atributos, TELÉMACO también permite especificar un *formato infijo de cadena*, el cual da un desarrollo simple).

Un *estilo de páginas* especifica todo el formato de instrucciones para un esquema de páginas dado; éste contiene un conjunto de atributos estilos para el esquema ADM de la página, además una *sección de encabezados* y una *sección de final (footer)*. El encabezado y final especifican las características gráficas que deben ser asociadas con la misma página, más exactamente con un atributo específico. Tanto los atributos estilos como los encabezados y los finales se componen de piezas arbitrarias de código HTML. Cuando se generan instancias de un esquema de páginas dado, PENÉLOPE carga el correspondiente estilo de página (para los esquemas de páginas que no tienen asociados un estilo de páginas, PENÉLOPE adopta un estilo por defecto) y formatos de datos según esto: cada esquema de páginas tiene definido el encabezado y el final en el estilo, y cada valor del atributo es encapsulado dentro de su formato de cadenas.

TELÉMACO hace estilos completamente transparentes para el diseñador y permite escribir páginas de ejemplos HTML, desde los cuales los estilos son producidos automáticamente; esas páginas HTML son llamadas *plantillas*. Una plantilla de páginas es un prototipo de páginas HTML que no contiene datos reales.

Los usuarios sólo trabajan con plantillas, y TELÉMACO tiene el cuidado de generar el correspondiente estilo de la siguiente manera:

- El punto inicial es una (o más) plantilla(s) de sitios, la cual puede ser deducida desde alguna página HTML existente, o creada de improviso, y progresivamente mejorarla hasta que la presentación sea satisfactoria.

- Cuando se lee la plantilla del sitio, se invoca TELÉMACO. El estilo de los sitios es utilizado otra vez por TELÉMACO para producir una primera versión de la plantilla de página, una para cada esquema de páginas en el comando PDL.
- Luego, son analizados las plantillas del sitio uno por uno (si es necesario), y se edita para mejorarlo.
- Cuando se completa la fase de edición, TELÉMACO puede ser invocado nuevamente para producir estilos de páginas desde las plantillas de página.

5. ***HOMER.*** Es una herramienta desarrollada para apoyar al diseñador a través de los pasos de diseño en el cual el diseño del sitio desarrolla distintos niveles y descripciones, cada uno basado en un modelo formal. HOMER posee dos facilidades: una interfaz gráfica de usuarios y un módulo que genera código automáticamente para ser ejecutados por las distintas herramientas y para implementar el sitio. Cuando se trabaja con HOMER, primero, el sistema toma como entrada una especificación declarativa del esquema conceptual inicial de datos y automáticamente lo traslada a un esquema lógico de base de datos (relacional); luego, su interfaz gráfica ayuda al diseñador a especificar transformaciones según las cuales se manipulan las construcciones del esquema conceptual de base de datos para obtener el hipertexto deseado y progresivamente se aplican esas transformaciones al diseñador de las formas del esquema ADM del sitio resultante. Una vez generada la descripción ADM para el sitio y basándose en las transformaciones especificadas, HOMER genera automáticamente el código PDL utilizado como entrada a la fase de creación de páginas.

6. ***NEPTUNE.*** Es un sistema que maneja el flujo de trabajo desarrollado para incorporarse con otras herramientas del sistema. En esta estructura, el desarrollo del sistema de información complejo se basa en un sitio hecho de varias porciones mezcladas entre sí:

   - Un catálogo de porciones, *páginas de acceso a los datos*, utilizado para acceso y presentar la base de datos del sitio.
   - Una o más porciones de *ejecución de flujo de trabajo* dando acceso a uno o más servicios a través de la ejecución de un flujo de trabajo.

   Toda la lógica del flujo de trabajo se maneja con NEPTUNE, el cual genera código Java para coordinar varias actividades; asignando tareas a actores y autenticando accesos al flujo de trabajo, si es necesario. El sitio es utilizado como una interfaz en el flujo de trabajo; es decir, toda la interacción entre actores y el sistema de manejo de flujo de trabajo se desarrolla a través de páginas en el sitio; esas páginas se generan vía cliente-servidor, donde NEPTUNE se encuentra en el lado del cliente y PENÉLOPE, al igual que TELÉMACO, en el lado del servidor, y contiene formas para colecciones de entradas a usuarios y ejecución de tareas. La comunicación entre el sitio y el sistema manejador de flujo de trabajo se basa en la base

de datos, el cual es utilizado para almacenar el estado de flujo de trabajo y entradas de usuarios.

7. ***Interfaz de Usuario.*** Se escribe completamente en HTML, así el usuario final y administradores pueden acceder al sistema desde algún cliente en la red. El sistema permite manipular datos provenientes de distintos sitios. Los sitios son divididos en *Sitios Externos*, es decir, sitios administrados por otros servidores, sobre el cual el WBMS no tiene control directo; y *Sitios Locales* o *Internos*, que son creados y administrados por el sistema, sobre los cuales se tiene control completo. Las operaciones permitidas en sitios externos son esencialmente consultas y sitios de almacenamiento local de datos. Además las consultas, actualizaciones y reestructuraciones también pueden ser realizadas en sitios locales.

   Para cada sitio, el sistema manipula datos de naturaleza heterogénea y esencialmente páginas HTML y tablas de base de datos. Con respecto a los sitios externos, las tablas de la base de datos contienen datos extraídos del sitio y localmente materializados para servir como fundamento para próximos cómputos. En el caso de sitios locales, las tablas son utilizadas para manejar datos que serán publicados en el sitio.

CONCLUSIONES

Los sistemas de Consulta y Modelado en la Web estudiados resaltan la importancia que tiene la construcción y el mantenimiento de sitios Web con grandes cantidades de datos y cuya concepción y transformación se volvería tediosa ante la variabilidad constante de la información. Estas herramientas están apoyadas en una fuerte metodología, consistente en la separación entre la estructuración, diseño de la presentación y el establecimiento de las rutas de navegación de las páginas cuyos datos se obtienen a través de lenguajes estructurados de consultas, semejantes en su sintaxis al SQL.

Los sistemas Araneus y Strudel son herramientas diseñadas para la construcción y mantenimiento de los sitios Web basados en Sistemas de Manejo de Base de Datos (DBMS). A través de este estudio se mostró la importancia que estas herramientas tendrán para la industria del *software* en cuanto a la separación de las actividades propias que conciernen a la creación y mantenimiento de los sitios Web, ya que determinan una división clara de las actividades que se deben desarrollar para lograr obtener un sitio dinámicamente, sin necesidad de que estas actividades se sobrepongan unas con otras.

Se pueden diseñar los esquemas de páginas en forma dinámica guardando en una base de datos toda la información que se desee desplegar en el respectivo sitio Web. Esto trae como ventaja que no es necesario tener que realizar mucho esfuerzo

de programación para modificar o actualizar el sitio generado, sino que mediante una aplicación de mantenimiento de bases de datos, la cual puede ser desarrollada bajo tecnología de programación Internet, se captura los nuevos datos que se van a mostrar y de forma casi inmediata el sitio Web queda actualizado sin tener que bajar y luego volver a montar las páginas HTML que sufrieron cambios.

Uno de los trabajos más importantes ahora radica en realizar un buen esquema de cada una de las plantillas correspondientes a las páginas que conformarán el sitio, ya que sus respectivos diseños no se pueden modificar en forma dinámica y al no tener que actualizarlos muy a menudo garantizarán lo que llamamos en Sistemas como «larga vida de productividad del sitio» (lo contrario sucede con aplicaciones mal diseñadas, en las cuales su ciclo de vida es corto y por lo tanto costoso). Por otra parte, pocas modificaciones en el diseño no generarán como resultado tener que bajar las páginas para cambiar el esquema de cada una de ellas, ya que si esto sucediera se perdería tiempo al volver a construir todo el sitio, no sin antes verificar que los nuevos cambios no hayan afectado la consistencia de todos los datos guardados en la base de datos, los cuales al final son los que generan el sitio Web; además temporalmente habría que bajar el sitio donde están las páginas mientras se vuelve a montar ya actualizado, por lo que en este tiempo ningún «internauta» podría acceder a él, y dependiendo de la naturaleza del negocio del sitio Web, ocasionaría hasta pérdidas de dinero.

Como consecuencia de todo lo descrito, el mantenimiento y en general la administración del sitio se hace más rápido y fácil por parte del administrador e inclusive, por ejemplo, si el sitio despliega información que es constantemente actualizada, se pueden crear niveles de permisos para que diferentes usuarios que despliegan sus datos en el mismo sitio puedan realizar ellos mismo las respectivas actualizaciones de las páginas, y el administrador puede encargarse de labores más importantes en la administración del servidor y, por lo tanto, del sitio.

GLOSARIO

- LENGUAJE DE CONSULTA: Trozo de DML que implica recuperación de información.

- MODELO DE DATOS: Colección de herramientas conceptuales para describir datos, relaciones de datos, semánticas de datos y restricciones de datos, que describen la estructura de una base de datos.

- DBMS *(Data Base Management System):* Sistema Manejador de Base de Datos. Consiste en una colección de datos interrelacionados y una colección de programas para acceder a esos datos. El objetivo principal de un DBMS es proporcionar un entorno en el que pueda almacenarse y recuperarse información de forma conveniente y eficiente.

- HTML *(HyperText Markup Language):* Lenguaje de marcas hipertextuales. Lenguaje de computadora empleado para especificar el contenido y el formato de un documento de hipermedios en *World Wide Web.* Es poco usual que los usuarios se encuentren con el HTMLL, ya que éste es un detalle interno.

- SQL *(Structured Query Language):* Lenguaje estructurado de consultas. Lenguaje de bases de datos relacional estándar.

- XML *(Extensible Markup Language):* Formato universal para documentos estructurados y datos en la Web.

Bibliografía

[1] ATZENI, P.; MECCA, G.; MERIALDO, P., Design and Maintenance of Data-Intensive Web Sites. Dipartimento di Informatica e Automazione, Universidad de Roma Tre y DIFA, Universidad de la Basilicata. Junio de 1997.

[2] D. QUASS, J.; WIDOM, R.; GOLDMAN, K.; HAAS, Q.; LUO, J.; MCHUGH, S.; NESTOROV, A.; RAJARAMAN, H.; RIVERO, S.; ABITEBOUL, J.; ULLMAN and WIENER, J., LORE: A Lightweight Object REpository for Semistructured Data. Proceedings of the ACM SIGMOD International Conference on Management of Data. Montreal, Canadá, junio de 1996.

[3] FERNÁNDEZ, M.; FLORESCU, D.; KANG, J.; LEVY, A.; SUCIU, D., Catching the Boat with Strudel: Experiences with a Web-Site Management System. AT&T Labs. Noviembre de 1997.

[4] FERNÁNDEZ, M.; FLORESCU, D.; LEVY, A.; SUCIU, D., Web-Site Management: The Strudel Approach.

[5] FLORESCU, D.; LEVY, A.; MENDELZON, A., Database Techniques for The World Wide Web: A Survey.

[6] FRATERNALI, P., Tools and Approaches for Developing Data-Intensive Web Applications: A Survey. Politécnico de Milano.