INTEGRANDO INFORMACIÓN DE FUENTES HETEROGENEAS ENFOQUES Y TENDENCIAS

RESUMEN
El crecimiento de las tecnologías Internet y de las fuentes de información dispersas en todos los asuntos, implica la tarea de obtener la información de un número extenso de fuentes disponibles en la Web. El sueño de poder consultar transparentemente la información proveniente de fuentes heterogéneas y geográficamente dispersas exige la puesta en práctica de estándares y la construcción de herramientas. En este artículo discutimos los aspectos de la integración de sistemas de la recuperación de datos.

PALABRAS CLAVES: Wrapper, Recuperación de información, Consultas Web.

ABSTRACT
The growth of internet technologies and dispersed information sources in all topics, imply the task of obtaining information from the vast number of information sources that are available on the World Wide Web (WWW). The dream of being able to consult transparently the information coming from heterogeneous and geographically dispersed sources demands the implementation of standards and the construction of tools. In This article we discuss the aspects of information retrieval systems integration.

KEYWORDS: *Wrapper, Information retrieval, Web Queries.*

NESTOR DARIO DUQUE MENDEZ
M.Sc en Ingeniería de sistemas
Profesor Asociado
Universidad Nacional de Colombia
Sede Manizales
ndduqueme@unal.edu.co

JULIO CESAR CHAVARRO PORRAS
Ph.D. (c) en Ingeniería. Área de énfasis: Ciencias de la computación
Profesor Asistente
Universidad Tecnológica de Pereira
jchavar@utp.edu.co

RICARDO MORENO LAVERDE
M.Sc. en Administración Económica y Financiera
Profesor Asistente
Universidad Tecnológica de Pereira
rmoreno@utp.edu.co

1. INTRODUCCIÓN

Con la aparición de la Web, surge un exceso de información y resulta complicada la manera de encontrar lo que se está buscando. Un mar de páginas para poder encontrar un grano de arena. Un problema común a nivel industrial hoy en día es la cantidad de información disponible en diferentes fuentes de datos heterogéneas, como Internet, librerías digitales, bases de datos antiguas, Sistemas de correo electrónico, etc. Las empresas en general, y los directivos en particular, necesitan disponer de esta información dispersa y heterogénea de una manera rápida, consolidada y relevante para poder tomar las decisiones adecuadas en cada momento.

Tradicionalmente las consultas se realizaban a una base de datos local o remota con un esquema bien conocido. En los sistemas de información actuales las consultas implican a diferentes fuentes de datos, que no sólo pueden estar dispersas geográficamente, sino que también suelen tener esquemas diferentes. Más aun, dicha información se encuentra como documentos en la Web.

La Web está basada en un paradigma de navegación lo que hace muy difícil la recuperación e integración de datos desde diferentes sitios. Generalmente para lograr este cometido se requiere de aplicaciones especializadas con dificultades de tiempo para desarrollo, y complicado y permanente mantenimiento. El crecimiento y los cambios de formatos de esta información son enormes en poco tiempo, lo que hace obsoletas las aplicaciones desarrolladas.

Lo ideal sería poder recuperar esta información en la Web como si estuviese en bases de datos; con la misma facilidad y transparencia de contenidos para el usuario. Por ejemplo:

Select vuelo, origen, destino, horario
From aerolinea
where fecha="XX/XX/XX"

Los trabajos sobre recuperación e integración de datos desde páginas Web para proporcionar información a usuarios, vienen tomando gran importancia en los últimos tiempos [1]. Este artículo muestra algunas de las técnicas utilizadas y algunas aplicaciones concretas.

2. WRAPPERS SOBRE FUENTES WEB

El constante desarrollo de forma creciente de Internet ha generado la necesidad de extraer la información contenida en las páginas Web de forma automática y con la mayor precisión posible. Los wrappers extraen información de páginas Web (u otras fuentes semiestructuradas) escritas en un formato específico.

Un wrapper sobre fuentes Web es considerado un software que acepta consultas de usuarios de datos en la Web y luego de extraer la información relevante retorna los resultados [2].

Fecha de Recepción: 31 Agosto de 2006
Fecha de Aceptación: 20 Marzo de 2007

Sus tareas se pueden resumir en:

- Aceptar consultas acerca de la información en las páginas fuentes
- Conseguir las paginas relevantes
- Extraer la información solicitada
- Retornar los resultados

En ocasiones el wrapper se encarga también de monitorear las fuentes de datos para detectar posibles cambios en los datos e informar de estos cambios al módulo de integración siendo este último componente el responsable de integrar los datos.

Algunas de las ventajas de construir wrappers en Web se pueden resumir en:

- Consultas en las fuentes similares a las realizadas sobre bases de datos
- Todas las fuentes sobre la que se construyó el wrapper pueden ser procesadas usando un lenguaje común de consulta.
- Este acceso integrado se puede hacer usando un mediador que integra la información desde las diversas fuentes

Pero también es necesario reconocer que son imprácticos en casos donde el número de fuentes de interés es muy grande, nuevas fuentes son adicionadas frecuentemente o se dan cambios usuales en el formato de fuentes.

2.1 Sistema Mediador-Wrapper

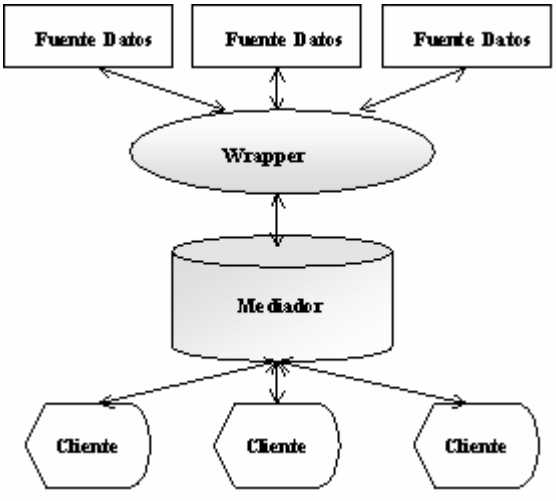

Figura 1. Arquitectura basada en mediador

Extendiendo la idea de las bodegas de datos, se puede proponer una arquitectura de mediador-wrappers, cada mediador maneja un esquema de datos común capaz de representar los datos ofrecidos por cada fuente de datos. Por cada fuente de datos existe un wrapper que conoce el esquema de datos de su correspondiente fuente de datos y traduce los datos al esquema de datos existente en el mediador.

Un mediador determina las fuentes a ser consultadas para responder una consulta y las sub-consultas a ser enviadas a cada fuente, optimizando los planes de ejecución de las consultas multifuentes. Como característica adicional, para el mediador es transparente el modelo de datos, lenguaje de interrogación y el esquema de datos, utilizado por cada fuente de datos.

El wrapper optimiza las consultas ha ser enviadas a una fuente y traduce las enviadas por el mediador hacia su fuente de datos.

Este sistema permite que los wrapper extraigan la información de fuentes heterogéneas y se integre a través de los mediadores. Los mediadores pueden ser bases de datos convencionales o bases de conocimiento representadas con datos estructurados extraídos desde fuentes semi o no estructurados [3]. Las consultas de los usuarios se aplican sobre estos almacenes de datos, en lugar de ir directamente a la Web [4].

2.2 Construcción de Wrappers.

La construcción de un wrapper incluye:

a. Identificar la estructura de las fuentes, identificando secciones y subsecciones de interés
b. Construir un parser o analizador para que trabaje sobre las páginas fuentes
c. Definir la comunicación entre wrapper, mediador y fuentes Web.

Cada fase se compone de ciertas tareas:

a. Identificar la estructura de las fuentes, identificando secciones y subsecciones de interés. La tarea de estructuración puede ser automática o con mínima interacción con el usuario a través de programas que analizan el documento HTML y luego otros que formatean la información con base en tokens (símbolos) y/o jerarquías. A la vez implica dos pasos:

i. Identificar los tokens (símbolos que indican o diferencian secciones) de interés en la página. Se usan analizadores léxicos que permiten obtener expresiones regulares en los documentos. Un ejemplo se ve en la tabla siguiente:

Regular Expression given as LEX Specification	*Description*	*Example Heading*
''<''[bB][^<>]*''>''[^\n]+''<''/[bB]''>''	Headings in bold tags	<a href=''''>Chair
''<''[hH][0-6][^<>]*''>''[^\n]+''<''/[hH][0-6]''>''	Headings with font size	Geography
''<STRONG>''[^\n]+''</STRONG>''	Headings in Strong Tags	<STRONG>Area</STRONG>
''<Strong>''[^\n]+ ''</Strong>''	Strong tags in different case	<Strong>Population</Strong>
''''[^\n]+''''	Strong tags in lower case	Deadlines
[A-Za-z0-9\-_]+[:]	Word sequence ending in colon	IRS NUMBER:
''<''[iI][^<>]*''>''[^\n]+''</''[iI]''>''	Italicized group of words	total area:

Tabla 1. Expresiones [6]

ii. Determinar las jerarquías dentro de las secciones. Generalmente se usan heurísticas basadas en el tamaño de la letra o la indentación de los documentos

b. Construir un parser o analizador para que trabaje sobre las páginas fuentes. A partir de los pasos anteriores y usando un analizador léxico y un generador de compilación, se leen las fuentes y se descubre la estructura. Ejemplos de estos productos son LEX[1] (LEXical analyzer Generator) y YACC[2] (Yet Another Compiler-Compiler). El primero divide la fuente de acuerdo a los tokens predefinidos y el segundo busca la estructura jerárquica.

c. Definir la comunicación entre wrapper, mediador y fuentes Web. Varios elementos se involucran en este paso:

- Determinar las URLs de las páginas de las cuales se va a extraer información, generalmente partiendo de la página inicio (index).
- Garantizar la capacidad de recuperar páginas Internet; generalmente con programas o script en Java, Perl, etc.
- Comunicación entre wrapper y mediador, usando un lenguaje de comunicación entre agentes como KQML u otro.

2.3 Tendencias en los Trabajos.

Buena parte de los trabajos actuales están enfocados hacia lograr la generación automática de los wrapper a partir de la planificación inteligente y la definición de reglas de aprendizaje en forma dinámica en las cuales se tenga como entrada reglas de entrenamiento, reglas admisibles de extracción y las reglas de búsqueda [4].

En algunos casos se han dado reglas de entrenamiento por ejemplo y prueba, con reglas admisibles para extracción basada en prefijos y sufijos de ítems de interés. Usando estrategias de búsqueda desde prefijos y sufijos cortos hasta ir encontrando los correctos. Se logra un rápido aprendizaje, pero se tiene problemas con los cambios o pérdida de datos y requiere un alto nivel de ejemplos.

El aprendizaje por transductores usa reglas contextuales, como: A la izquierda se tienen letras MAYUSCULAS, y la derecha se termina con , etc. En estos casos se lee primero para encontrar los límites entre tokens y en una segunda lectura se buscan patrones contextuales.

Los wrapper por inducción jerárquica, *wrapper Stalker* [8], como se muestra en la figura 2, resuelven un problema grande descomponiéndolo en otros más pequeños y atacándolos en forma individual, haciendo la extracción independientemente en cada uno.

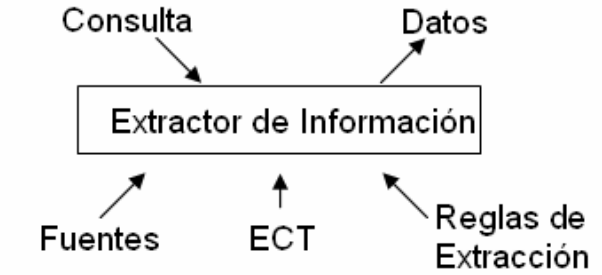

Figura 2. Wrapper por inducción jerárquica

La fuentes son exploradas con base en reglas de extracción (vistas como un secuencia de landmarks o fronteras) y un árbol de catalogo embebido (ECT) donde se define la estructura arbórea de la fuente. La información así recuperada permite devolver la respuesta a la consulta del usuario.

3. APLICACIONES

Siguiendo la línea en [2], se presentan algunos de los proyectos más reconocidos en este campo:

ARIADNA: El proyecto Ariadna [5] se centra en el desarrollo de metodologías y herramientas que construyen de forma rápida agentes inteligentes, similares a los wrappers, con el objetivo de unificar vistas de recursos Web. Las herramientas permiten la construcción de wrappers que hacen búsquedas y consultas en fuentes WWW como si se tratara de una base de datos tradicional. Su estrategia es descomponer consultas y combinar las respuestas.

Ariadna integra datos de sitios múltiples, integrando datos fuentes Web semi-estructuradas. Crea agentes de información y da una visión unificada a diversos recursos

[1] http://www.ds9a.nl/lex-yacc/cvs/output/lexyacc.html

[2] http://dinosaur.compilertools.net

Web. Permite consultas tipo base de datos convencional, descompone las consultas y combina las respuestas.

Componentes:

- Un modelo de dominio del problema
- Una descripción de las fuentes de información en términos del modelo.
- Wrapper que proveen acceso uniforme a las fuentes de información permitiendo consultas como si fueran bases de datos relacionales
- Un planificador de consultas que dinámicamente determina la forma eficiente de procesar una consulta de usuario de acuerdo a las fuentes de información disponibles. El sistema descompone la consulta en subconsultas a fuentes de datos individuales.

Desde esta visión se tratan las páginas Web como una fuente de información que puede ser consultada, por lo cual requiere de un wrapper que pueda extraer y retornar la información solicitada desde cada tipo de página.

La generación semi-automática de un wrapper para una determinada fuente se lleva a cabo en los siguientes pasos:
- Estructurar la fuente: implica identificar las secciones y subsecciones de interés de una determinada página.
- Construcción de un parser: se encargará de analizar la estructura derivada en el primer paso.
- Añadir capacidades de comunicación entre el wrapper, el integrador y las fuentes WWW.

Los wrapper de Ariadna poseen:
- Un modelo Semántico
- Un modelo Sintáctico
- Una interfase de usuario orientada a demostración (Demonstration-oriented user interface DoUI), donde el usuario muestra al sistema que información extraer desde páginas ejemplos. Bajo la interfaz está un sistema de máquina de aprendizaje para inducir las reglas gramaticales.

TSIMMIS: Acrónimo de *The Stanford-IBM Manager of Multiple Information Sources* Como su nombre relacionado en Yiddish, es una mezcla deliciosa de cosas diversas. La meta del proyecto de TSIMMIS es desarrollar herramientas que faciliten la integración rápida de fuentes de información heterogéneas que pueden incluir datos estructurados y semiestructurados. Sus componentes pueden traducir preguntas e información, extraer datos de la Web y combinar información de diversas fuentes vía mediador [7].

World Wide Web Wrapper Factory (W4F) [9]: Es un proyecto que genera wrappers bajo Java. Su proceso se identifica con las tres siguientes etapas: Recuperación, extracción y traslación o mapeo. El lenguaje de recuperación especifica dónde y cómo cargar el documento HTML. Una vez recuperado el documento se convierte en un árbol y se aplican las reglas de extracción según los datos que se quiera conseguir del documento. La información extraída es almacenada en un formato interno basado en listas anidadas. Finalmente estas listas son trasladadas al formato final para su presentación al usuario. Este proceso es repetido para cada documento Web. El proyecto W4F incluye asistentes para cada uno de las fases antes descritas que nos ayudan a la hora de especificar los datos que se quiere extraer de una determinada fuente.

TUKWILA: Es un sistema de integración de datos que trata de responder a las consultas realizadas a diferentes fuentes de datos autónomas y heterogéneas. Todas las fuentes son trasladadas a un esquema intermedio común. El sistema de integración de datos reformula las consultas en varias consultas sobre las fuentes y posteriormente combina el resultado de todas en una sola respuesta. El objetivo de Tukwila es procesar consultas eficientemente de cadenas de datos en XML.

4. CONCLUSIONES

La utilización de la creciente acumulación de información a la vez que dispersa y disímil se convierte en reto ineludible en el mundo actual. La complejidad asociada a esta tarea es innegable. Las consultas integradoras en tiempo real, requiere aún otros esfuerzos.

La combinación de diferentes técnicas y la aplicación de variantes exitosas en otros campos se imponen. Una de las posibilidades se basa en los agentes de recuperación de información que representen los wrappers en sus diferentes variantes. Los agentes de recuperación de información pueden poseer métodos que permitan el rápido acceso y recuperación de información relevante, administrando, manipulando y juntando información de fuentes distribuidas, entrenados con mecanismos de búsqueda y navegación flexibles y algoritmos de clasificación poderosos.

5. BIBLIOGRAFÍA

[1] MUSLEA, I., MINTON, S., & KNOBLOCK, C. AHierarchical wrapper induction for semistructured information sources. Autonomous Agents and Multi-Agent Systems. 2001.

[2] ARAQUE CUENCA, HURTADO TORRES. SAMOS JIMENEZ. Extracción de Información de Fuentes de Datos Heterogéneas e Incorporación al Data Warehouse. Universidad de Granada

[3] Inmon, W.: "The Operational Data Store. Prism Tech Topics. Vol.1, No.17. 1993.

[4] ARAQUE CUENCA. Extracción de Información de Fuentes de Datos Heterogéneas con características Temporales. Universidad de Granada

[5] KNOBLOCK KNOBLOCK, C. et. al. The Aradne Approach To Web-Based Information Integration. 2004.

[6] ASHISH NAVEEN, KNOBLOCK CRAIG. Wrapper Generation for Semi-structured Internet Sources. The University of Southern California Information Sciences Institute. http://www.isi.edu/info-agents/papers/ashish97-ssd.pdf

[7] LERMAN, K. MINTON, S., KNOBLOCK, C. Wrapper Maintenance: A Machine Learning Approach. USC Information Sciences Institute. 2003. Disponible en: http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume18/lerman03a.pdf.

[8] CHEN LI, RAMANA YERNENI, VASILIS VASSALOS, HECTOR GARCIA-MOLINA, YANNIS PAPAKONSTANTINOU, JEFFREY ULLMAN, MURTY VALIVETI. Capability Based Mediation in TSIMMIS. SIGMOD 98 Demo, Seattle, Junio 1998.

[9]. BHANDARI, DEEPALI. Extraction Of Web Information Sing W4f Wrapper Factory And Xml-Ql Query Language. Philadelphia, Pennsylvania. 1999.