Generación Automática de una Base de Datos desde Documentos de la Web

Jaime Ferreiro, Regina Motz , Fernando Perelló, and Dina Wonsever

Universidad de la República, Montevideo, Uruguay

e-mail: [rmotz,wonsever]@fing.edu.uy, fperello@movinet.com.uy,jotaefe@adinet.com.uy

Resumen

El objetivo central de este trabajo es la extracción de información de documentos HTML y la consolidación de esta información en una base de datos. Se propone un mecanismo basado en una ontología del dominio, en patrones sintácticos típicos para la inferencia de algunos tipos de datos y en heurísticas para la interpretación de títulos y tablas. Mediante este mecanismo se construye automáticamente un mapeo entre elementos de un documento HTML y las entidades del dominio del usuario. Utilizando este mapeo, se transfiere la información extraída de la página Web a una base de datos local.

1 Introducción

La *World Wide Web* (WWW) ha crecido vertiginosamente en los últimos años y es actualmente una fuente fundamental de información en prácticamente todas las áreas de interés. Los sistemas de manejo de la información global permiten a los usuarios clasificar los documentos y acceder a versiones completas de los mismos. Sin embargo, si se manejan requerimientos que implican un manejo integrado de la información, tales como encontrar el computador más barato de las ofertas *online* existentes o construir una tabla de los hoteles ofrecidos en cada ciudad de una región, los usuarios necesitan extraer, sintetizar y mantener información de diversos documentos. Estas tareas requieren un considerable esfuerzo.

El objetivo de este trabajo es proveer asistencia en la realización de las tareas mencionadas. En este sentido, se propone un mecanismo para interpretar de modo automático información semiestructurada de documentos HTML pertenecientes a un dominio específico.

El mecanismo propuesto extrae información del dominio a partir de los documentos HTML y la almacena en una base de datos local, que puede ser consultada por usuarios y cuya información puede seguir siendo procesada. Se utiliza un esquema orientado a objetos para representar los requerimientos del usuario y una ontología para describir el dominio de interés. El proceso de extracción de información es automático y se apoya en la interpretación de elementos tales como tablas y títulos, guiándose por la ontología del dominio y los requerimientos del usuario.

Apoyándose en técnicas ya existentes en el área de Bases de Datos, tales como las propuestas en [HGMC+97, HFAN98] se propone un mecanismo basado en una ontología del dominio, en patrones sintácticos típicos para la inferencia de algunos tipos de datos y en heurísticas para la interpretación de títulos y tablas. Mediante este mecanismo se construye un mapeo entre elementos

de un documento HTML y las entidades del dominio del usuario. Utilizando este mapeo, se transfiere la información extraída de la página Web a la base de datos local.

Se detalla a continuación la organización del resto del artículo. En la sección 2 se presenta un análisis de los distintos tipos de información que es habitual encontrar en una página Web, y de las dificultades que plantea la explotación de los mismos. La sección 3 está dedicada a presentar de modo general la propuesta de extracción de información realizada, que se detalla según sus fases en la sección 4 (análisis sintáctico) y en la sección 5 (análisis semántico). Finalmente (sección 6) se presentan conclusiones y posibles extensiones.

2 Tipos de información en una página Web

En una página Web se encuentran datos de distintos tipos y con diversa semántica. En particular, podemos encontrar imágenes, sonidos (menos habitual) y texto con distintos niveles de estructuración. Dentro de la información tipo texto encontramos texto libre, títulos, tablas, listas, vínculos. En este trabajo nos hemos concentrado en la explotación de algunos casos de información tipo texto, en particular títulos y tablas. La hipótesis es que mucha información relevante y traducible al formato de una base de datos aparece en estos tipos de datos. Analizamos a continuación cómo aparecen títulos y tablas en documentos HTML y las dificultades que se presentaron para extraer información de dichos elementos :

Títulos
Es importante explotar la información contenida en los títulos de un documento HTML, ya que suelen contener datos que se desea resaltar. A nivel del código HTML, si bien hay *tags* específicos para la definición de títulos, éstos rara vez son utilizados. En la práctica se logra los efectos de resalte deseados a través del uso de tags font (tipo de letra) y de especificación de propiedades como tamaño o color.
Además, el uso de tags específicos permitiría el rápido reconocimiento de título y subtítulos. Ante la ausencia de estos tags, la jerarquía de títulos debe ser inferida.
Es de notar que en este trabajo sólo se analizan los títulos próximos a tablas, con el objetivo de interpretar adecuadamente los datos de las mismas.

Tablas
Entendemos por tabla la información que es visualizada (mostrada por el browser) en forma tabular, es decir, en filas y columnas conteniendo datos en celdas. Estas tablas son bastante frecuentes en la Web y son implementadas mediante los tags <TABLE>, <TH>, <TR> y <TD> conjuntamente con sus respectivos atributos.
Los problemas más importantes encontrados al tratar de extraer información de tablas son :

- Muchas veces en HTML se utilizan tablas con propósitos de diseño y no para desplegar información. Se plantea entonces el desafío de discernir cuando una tabla es utilizada con propósitos de diseño y cuando es utilizada para mostrar información.
- Otro problema que se plantea es que es frecuente encontrar en Internet páginas donde los datos de una celda contienen múltiples datos, por ejemplo, el nombre y la dirección, o incluso, otra

tabla entera (tablas anidadas). Para los propósitos de este trabajo sólo se consideran tablas que contienen un único dato por celda.

- Para interpretar los datos de una tabla es necesario conocer los conceptos asociados a las columnas. Para esto es necesario, o bien, identificar una línea de títulos de columna o sino deducir a partir de los propios datos el concepto asociado. Uno de los aportes de este trabajo es en este sentido.

Además de los problemas enumerados asociados a cada tipo de datos en la Web encontramos problemas generales vinculados al uso del código HTML.
Estos problemas derivan de las inconsistencias dentro del código HTML, las cuales pueden ocasionar errores al reconocer tags dentro de una página. Es frecuente encontrar marcas que no tienen su homóloga o que tienen un alcance traslapado y a pesar de ello pueden ser interpretadas sin problemas por un navegador. Estos problemas fueron resueltos con la herramienta Tidy, recomendada por el World Wide Web Consortium [W3C].

3 Arquitectura del sistema propuesto

El sistema propuesto realiza las siguientes tareas : *reconocer, clasificar y consolidar* la información que se encuentra en forma de tablas y sus títulos encontrada en las paginas Web analizadas.

Reconocer: Implica fundamentalmente aislar las porciones de una página Web que van a ser sometidas a un análisis detallado. Se han definido restricciones para los elementos que se analizan y heurísticos para la determinación de títulos e ítems.

Clasificar: Implica inferir los conceptos de los datos que son encontrados y reconocidos dentro de las restricciones definidas para la extracción de la información. Se utilizan heurísticos y una ontología del dominio.

Consolidar: Implica realizar la carga efectiva de los datos extraídos en una base de datos. Se tiene en cuenta la especificación de la realidad que el usuario necesita y que está dada por una representación en esquema ODMG.

A continuación (Figura 1) se especifican la estructura del sistema y las fases de procesamiento.

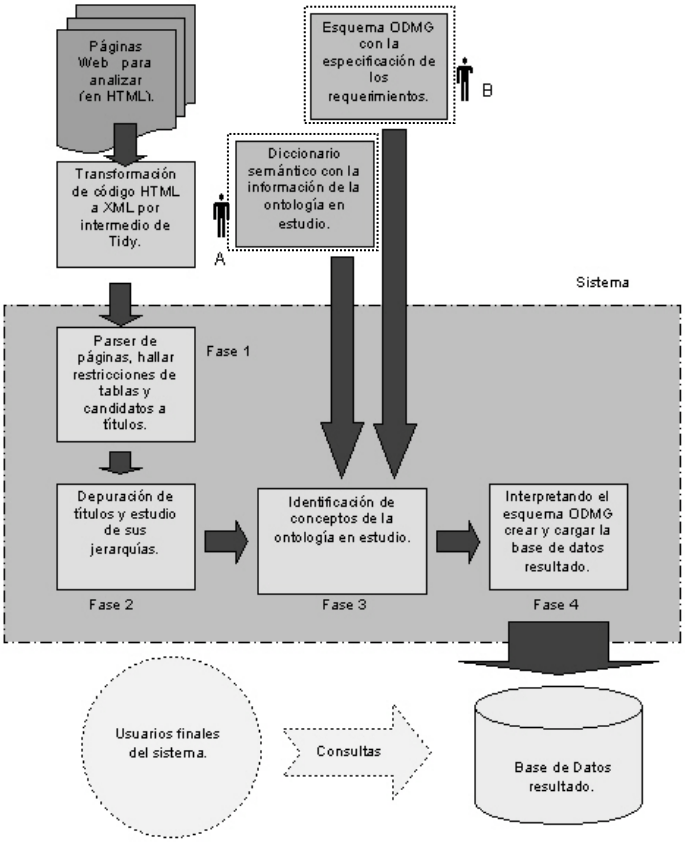

Figura 1: Arquitectura del Sistema

Los rectángulos con línea punteada marcados con el dibujo de usuario y nombrados como A y B, nos indican las entradas que dependen de los usuarios. Distinguimos dos tipos de usuarios: Super Usuario (A) y Usuario Modelador (B). El Super Usuario es el encargado de definir los conceptos del dominio elegido, es decir, es el que nos proporciona el diccionario semántico para nuestro análisis. El Usuario Modelador cumple el rol de definir los requerimientos : modelo de datos. Este modelo nos indica que información tenemos que buscar dentro de las paginas en estudio y será especificado con un esquema ODMG.

Las páginas Web son aquellas que contienen la información que el usuario del sistema desea conocer. Las páginas que se analizan están dadas explícitamente por el usuario (es un conjunto de páginas o direcciones finito y conocido). Un punto importante a tener presente (representado en el diagrama) es que si bien las páginas originalmente están en HTML se transforman a XML por intermedio de Tidy [Tidy00].

El dibujo de la base de datos resultado en el esquema, representa la información obtenida luego de la aplicación del sistema a las paginas Web y tiene la estructura especificada por el esquema ODMG.

El procesamiento que se realiza sobre las páginas se ha estructurado en fases. A continuación se caracteriza de modo resumido cada fase :

Fase I : Esta fase es la encargada de leer los códigos de las páginas recibidos en XML (aplicándoles un parser), encontrar las tablas que cumplen las restricciones y los candidatos a títulos mediante heurísticas (ver sección 4) definidas para este propósito.

Fase II: La segunda fase se encarga de descubrir las jerarquías del documento [Riloff93] y de eliminar los títulos que no parecen corresponderse con la jerarquía de títulos encontrada. Para ello utiliza heurísticos de cálculo de pesos y un heurístico de depuración de títulos.

Fase III: A partir de esta fase de la solución del problema los heurísticos comienzan a estar ligados al diccionario semántico que instancia la ontología considerada. El trabajo tiene 2 etapas : en primer lugar, identificar los conceptos y valores de los títulos obteniendo así objetos semánticos (ver sección 5); en segundo lugar, identificar los conceptos de cada una de las columnas de las tablas con lo que podremos obtener objetos semánticos considerando el texto de cada una de las celdas como el valor asociado al concepto de la columna correspondiente.

Fase IV: Esta fase es la encargada de crear la instancia de Base de Datos para el ODMG dado. Recibe de fase III los objetos semánticos complejos de ODMG que pudieron ser identificados con la ontología (ver sección 5)

4 Análisis sintáctico

En esta etapa se realiza un análisis de formato para identificar títulos, jerarquía de los mismos y columnas de las tablas. Se trabaja sobre la estructura construida por un parser que opera sobre el resultado de Tidy (página Web de entrada en formato XML). Contando con dicha estructura y con un conjunto de heurísticos surgidos de la investigación en la Web de las distintas formas de encontrar información, se inicia un análisis y depuración del código para terminar obteniendo una versión del documento *parseado* en la cual está la información que podría llegar a ser relevante para satisfacer los requerimientos (el esquema ODMG). Los heurísticos también se encargan de resolver los niveles de jerarquía encontrados dentro de la página.

Se especifica en primer lugar restricciones que deben satisfacer las tablas y títulos sobre los que se trabaja y luego las heurísticas utilizadas.

Restricciones sobre tablas

Se definen restricciones de tabla para separar los casos de utilización de los tags de tablas que contienen realmente información en forma tabular de otros que no interesa analizar (p.ej., tags de tabla utilizados con propósitos de diseño).

Se reconocen como tablas porciones de código que :

- Estén contenidas entre los tags <TABLE> y </TABLE>
- Presenten una estructura repetitiva de información tipo texto. Para cada fila (tag <TR>) se cumple que todas tienen igual número de celdas (igual número de <TD> </TD>), las celdas sólo tienen atributos o tags de texto y alguna celda de la fila debe contener información .

Restricciones sobre títulos

Se definen restricciones para identificar porciones de una página Web que son candidatos a título. Estos candidatos serán posteriormente validados o no por una heurística definida a esos efectos.
Las consideraciones que tomamos en las restricciones de inferencia de títulos de tabla son las siguientes:

- Serán candidatos todos aquellos textos que se encuentren dentro de los tags específicos de títulos o de resalte o de centrado: <P>,<CENTER >, <B>, <I>,<H1-6>,etc.
- Los títulos serán considerados como unidades indivisibles, es decir, si se encontraran dos títulos seguidos los consideraremos como dos distintos y no los dos juntos formando un solo título
- La longitud del título debe ser menor que un cierto número de caracteres.

Los heurísticos

A continuación, enunciamos y explicamos los heurísticos utilizados:
Heurístico de reconocimiento de títulos: luego de hallar una tabla que cumpla las restricciones, el siguiente paso será hallar los posibles candidatos a titulo de la misma. Para ello, se buscan sucesivos candidatos a título, desde el comienzo de la tabla hacia el comienzo del documento.

Heurístico para encontrar conceptos y valores en los títulos: Se buscan ocurrencias de los términos de la ontología (términos principales, sinónimos y valores) dentro del texto del candidato a título

Heurísticos para hallar el concepto de una columna: Las celdas de una columna de una tablas tienen por lo general la información correspondiente a un concepto. Esta puede presentarse con un cabezal en la primera celda o sin cabezal con los datos únicamente. Para identificar el concepto se utilizan dos vías : ubicación en la ontología del título de columna o (si ésta falla) inferencia a partir de expresiones regulares que identifican algunos tipos de datos usuales (nombres propios, direcciones, teléfonos, etc.)

Heurístico de cálculo de peso para títulos: Para cada candidato a título se calcula un peso para poder inferir el árbol de jerarquía del documento. En el cálculo del peso intervienen el tipo de tag y los atributos tipo de letra y tamaño.
A continuación mostramos una tabla donde aparecen los tags de texto utilizados y los pesos asignados.

TAG	PESO ASIGNADO	DESCRIPCIÓN
<h1>	30	Título de mayor jerarquía
<h2>	25	Título de segunda jerarquía
<h3>	20	Título de tercera jerarquía
<h4>	15	Título de cuarta jerarquía
<h5>	10	Título de quinta jerarquía
<h6>	5	Título de sexta jerarquía
<u>	3	Texto subrayado
<p>	0	Párrafo
<b>	2	Letra negrita

<I>	1	Letra itálica
	0	Utilizado para alinear texto
<font>	3	Define el tamaño, color y tipo de letra
<a>	0	Un link
	0	Centrado

Figura 2 : Pesos para jerarquía de títulos

Heurístico para la depuración de títulos: Se basa en el resultado del heurístico anterior y su función es determinar si un candidato a título es realmente un título o no. Para hacer esto, se compara los pesos obtenidos por el heurístico de pesos.
Se apoya sobre la siguiente hipótesis: No puede existir un título de menor jerarquía (peso) seguido de uno de mayor jerarquía. Los títulos de la tabla serán los candidatos a título más cercanos a ella que cumplan la anterior afirmación. Ver figura 3

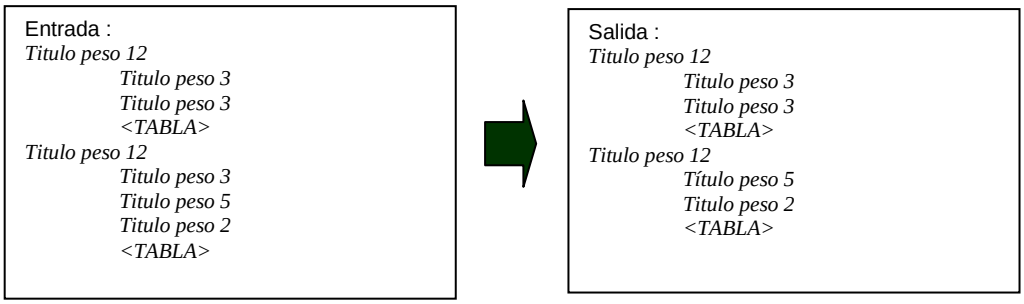

Figura 3 : Ejemplo de depuración de títulos.

5 Análisis semántico

Luego de obtener la información relevante, resultado de la fase descripta en la sección anterior, es necesario descubrir cual es la semántica de la información encontrada en esa página. Para esto utilizamos el diccionario semántico que define una ontología. Los conceptos especificados por el usuario a través del modelo de datos (en nuestro caso ODMG [CB97]) son identificados en la ontología y los datos correspondientes a estos conceptos son extraídos de la página.

Para asegurar la correcta interpretación de los datos extraídos es necesario trabajar sobre un dominio específico. En cada dominio existen conceptos básicos a los cuales llamaremos *conceptos de ontología*. Por ejemplo dentro del dominio específico *hotel*, los conceptos de ontología son: *ubicación, servicios, cuarto,* etc.

Un concepto de ontología puede abstraerse a un conjunto de palabras, abreviaciones, sinónimos y formatos característicos.

La palabra ontología genera controversia [Gruber]. En filosofía se refiere al sujeto de existencia, en términos computacionales una ontología es una especificación de una conceptualización [Gruber95, Riloff93, Hwang99]. En este trabajo es usada como una descripción de conceptos y sus relaciones. Podemos entonces modelar la ontología con el MER de la Figura 4.

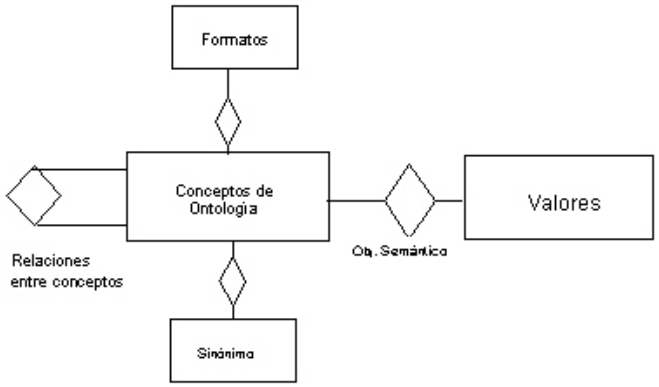

Fig. 4: Modelo Entidad Relación para una Ontología.

En la entidad ***conceptos de ontología*** están los conceptos definidos para el área específica. En ***formatos*** se almacenan las expresiones regulares en las que se presentan las instancias de los conceptos de ontología. Por ejemplo, para un número de teléfono aparece la expresión regular xxx-xx-xx donde x es un dígito decimal. En ***sinónimo*** se encuentran sinónimos propiamente dichos del concepto de ontología (por ejemplo hostería en lugar de hotel) y abreviaciones (por ejemplo Dpto. en lugar de Departamento). En ***valores*** aparecen las palabras que se relacionan con un concepto de Ontología. Un ejemplo de esto es el caso de los nombres de departamentos para el concepto departamento. Los sinónimos o abreviaciones de un valor de un concepto se encuentran como un valor distinto, por ejemplo aparecen Montevideo y Mdeo. La relación ***relaciones entre conceptos*** es utilizada para modelar las inclusiones y pertenencias entre conceptos, por ejemplo el concepto ciudad esta incluido en el concepto departamento.

Dentro de una ontología se distinguen objetos semánticos simples y complejos [Bornhövd99].

Un ***objeto semántico simple*** es una tripla <C,v,$>, donde C es el concepto de ontología al cual pertenece, v el valor del objeto semántico y $ el contexto utilizado para interpretar correctamente. Un ejemplo de objeto semántico es <distancia,10,"Km">.

Un ***objeto semántico complejo*** es una colección heterogénea de objetos semánticos simples. Se representa también por una tupla <C,A> donde C es el concepto de ontología y A es un conjunto de objetos semánticos simples. A esta dividido en dos subconjuntos $\underline{A}$ y A_R. $\underline{A}$ es el conjunto de atributos clave que son usados para identificar a un objeto semántico complejo del concepto C. Estos atributos son determinados por C y especificados en la ontología. El subconjunto A_R representa atributos adicionales. A_R puede variar entre diferentes objetos semánticos del mismo concepto de ontología.

La información que el usuario desea extraer de las páginas es especificada usando el modelo de datos ODMG. Elegimos ODMG por ser el estándar para modelo de datos orientados a objetos y por su riqueza semántica para modelar estructuras.

La correspondencia entre las clases y atributos ODMG y la información de las páginas se realiza en varios pasos.

El primer paso consiste en encontrar en la ontología los conceptos especificados en el modelo ODMG. Por ejemplo si el modelo ODMG define una clase de nombre Hostería, es necesario identificar en la ontología a qué concepto nos estamos refiriendo. En primera instancia, se utiliza el propio nombre de la clase, que se trata de mapear a un concepto con el mismo nombre. Si no existe un concepto con el mismo nombre, se llega a un concepto de la ontología a través de sus sinónimos. El mismo proceso se realiza para cada uno de los atributos de Hostería como puede ser nombre, dirección, teléfono, fax, etc. Se deduce además cuales son los conceptos necesarios para poder identificar a cada objeto de las clases, por ejemplo para identificar a un hotel u hostería es necesario conocer su nombre y su dirección.

El segundo paso consiste en identificar los objetos semánticos simples de las páginas Web guiados por los conceptos de la ontología obtenidos a partir del esquema ODMG. En el ejemplo de la figura 7, los objetos semánticos simples identificados son *departamento, ciudad, nombre, dirección* y *teléfono*.

Por último cada objeto semántico complejo es comparado con las estructuras guardadas del ODMG, en caso de ser identificado con alguna de ellas corresponde a un requerimiento y en este caso se extraen los valores que se insertan en la base de datos.

En la figura 7 el objeto semántico complejo resulta entonces Hotel, formado por el conjunto de objetos simples ya mencionados.

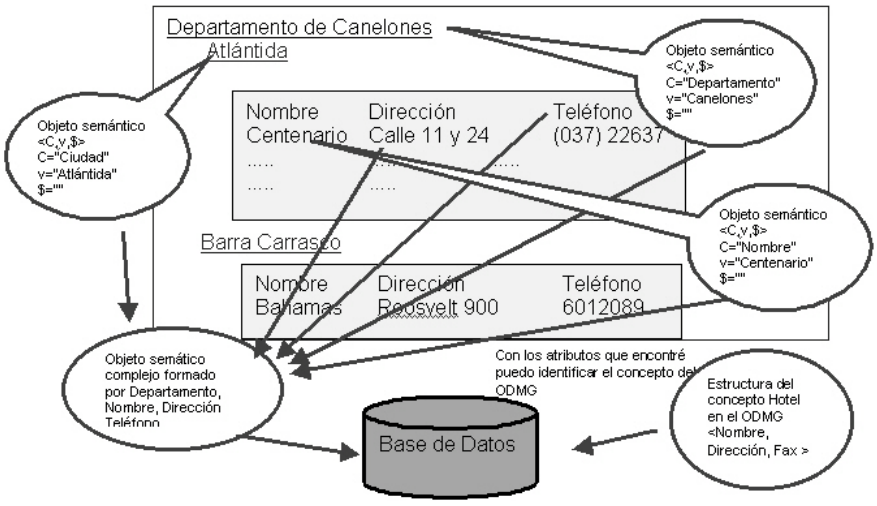

Fig. 7 : Correspondencia ente conceptos y datos de una página

6 Conclusiones

La principal contribución de este artículo es la propuesta de un mecanismo para la generación automática de una base de datos local que describe un dominio especificado por el usuario instanciado desde documentos HTML.

El proceso de extracción de los datos manipula en particular tablas que pueden o no contener títulos en sus columnas. Para realizar este análisis utilizamos una ontología que soporta la equivalencia semántica entre conceptos y que permite definir reglas de identificación de dominios, como pueden ser por ejemplo, teléfonos, e-mails o direcciones.

Otro punto importante para resaltar es que el mecanismo es capaz de identificar los distintos niveles de títulos en base no solo a los tags de títulos sino principalmente en base a su semántica. Así resulta posible resolver automáticamente el mapeo entre elementos del modelo de datos propuesto por el usuario (metadatos) y lista de ítems de los documentos HTML (datos).

Construimos un primer prototipo [FP00] implementado en Java. Para un conjunto de evaluaciones realizadas la generación de la base de datos resultó satisfactoria, sin embargo el conjunto de heurísticas propuestas necesita ser testeado con un número mayor de páginas. Como posible extensión se propone también el análisis y explotación de la información contenida en los vínculos, incorporando eventualmente las páginas a las que estos refieren.

Bibliografía

[AK97] Naveen Ashish, Craig Knoblock - Wrapper Generation for Semi-structured Internet Sources – Proc. Workshop on Management of Semi-structured Data, Tucson, Arisona, 1997

[AMM97] Paolo Atzeni, Giansalvatore Mecca, Paolo Merialdo, Semistructured and Structured Data in the Web: Going Back and Forth. Proceedings of ACM SIGMOD Workshop on Management of Semi-structured Data 1-9

[Bornhövd99] Christof Bornhövd Semantic Metadata for the Integration of Web-based Data for Electronic Commerce – Proceedings Int. Workshop on e-commerce and Web-based Information Systems, Santa Clara, 1999

[CB97] R. Cattell, D. Barry, “Object database standard : ODMG 2.0”, Morgan Kaufmann, 1997.

[CM97] Mary Elaine Califf and Raymond J. Mooney Relational Learning of Pattern-Match Rules for Information Extraction - Working Papers of the ACL-97 Workshop in Natural Language Learning 9-15

[DEW97] Robert B. Doorenbos, Oren Etzioni, and Daniel S. Weld - A Scalable Comparison-Shopping Agent for the World-Wide Web - Department of Computer Science and Engineering University of Washington – Agents 97 Conference

[FP00] J. Ferreiro , F. Perelló, “Extracción de Estructuras para Consolidación de Datos de la Web”. Proyecto final de la cerrera Ingeniería en Computación, Facultad de Ingeniería, UdelaR, Uruguay. Abril 2000

[Gruber] Tom Gruber What is an Ontology ? http://www-ksl.standford.edu/kst/what-is-an-ontology.html

[Gruber95] Thomas R. Gruber Toward Principles for the Design of Ontologies Used for Knowledge Sharing – International Journal of Human and Computer Studies 43(5/6), 1995

[HFAN98] G. Huck, P. Fankhauser, K. Aberer, E. Neuhold, "JEDI: Extracting and Synthesizing Information from the Web", COOPIS 98, New York, August, 1998. IEEE Computer Society Press.

[HGMC+97] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, A. Crespo, "Extracting Semistructured Information from the Web", Proc. of the Workshop on Management of Semistructured Data. Tucson, Arizona, May 1997.

[Hwang99] Incompletely and Imprecisely Speaking : Using Dynamic Ontologies for Representing and Retrieving Information. KRDB 1999 14-20

[NAM98] Svetlozar Nestorov, Serge Abiteboul, Rajeev Motwani – Extracting schema from semistructured data. In Proc. 11th International Conference on Data Engineering pages 295-306

[Riloff93] Ellen Riloff – Automatically Constructing a Dictionary for Information Extraction Tasks. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 811-816

[Soderland97] Stephen Soderland – Learning to Extract Text-based Information from the World Wide Web. Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97)

[Tidy00] Tidy http://www.w3.org/People/Raggett/tidy/

[Vier 99] D. Viera. "Extracción y Mantenimiento Dinámico de Datos de la Web". Proyecto final de la cerrera Ingeniería en Computación, Facultad de Ingeniería, UdelaR, Uruguay. Abril 1999.

[W3C] World Wide Web Consortium www.w3c.org