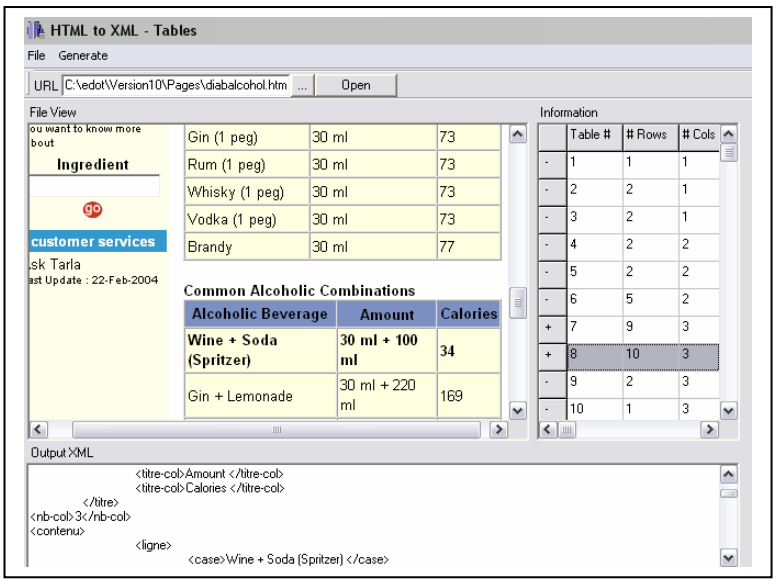
Extracción Semi-automática de Información a partir de tablas HTML



Marco-Javier SUAREZ-BARON.
Ingeniero de Sistemas.
Profesor auxiliar Facultad de Ingeniería UPTC Tunja.
Residences Fleming Paris-France,
Teléfono: ++ 33 1 69 15 34 76.

RESUMEN

Este trabajo presenta un proceso de extracción e integración de información contenida en tablas de tipo HTML, el proceso de extracción de información se apoya en un conjunto de reglas heurísticas y una clasificación rigurosa de tablas según su tipo de estructura y complejidad.
El propósito de la investigación es generar Documentos de tipo XML para integrar la información, obtenida en el proceso de extracción.

PALABRAS CLAVE

Extracción de Información, reglas Heurísticas, etiquetas, información semi-estructurada.

KEYWORDS

Semi-structured data, HTML, XML, DTD.

ABSTRACT

In this paper, I present an approach for information extraction from html tables that belong to a differents domains of interest.
This process is based in a detecting table-specific elements and the classification of html tables given their structure and distribution of information.

In my extraction approach, I capture each data source as semi-structured data, the result of information extracting of different HTML tables is presented in an XML document.

1. INTRODUCCION

Como la Internet y las Tecnologías Web están dirigiendo el curso de las diferentes formas de presentación y acceso a los datos y además es percibido como una red global de fuentes de datos heterogénea, es importante el desarrollo de aplicaciones informáticas que permitan la integración de estas fuentes de información.
Una de estas formas de presentación de datos e información son los documentos HTML, y en términos más específicos las tablas HTML, que
Con su aparición revolucionaron el diseño de las páginas Web, para mejorar la presentación y comprensión de datos e información. A pesar de que muchas de estas tablas HTML puedan tener una estructura única (si existiera un estándar para ello), es posible encontrar diferentes formatos para la distribución y organización de los datos dentro de la misma.
Para lograr una integración de la información contenida dentro de tablas HTML, es necesario, determinar el dominio o área de conocimiento (Salud, Turismo, Nutrición, etc.), una clasificación de tipos de tablas, y el formato de salida para la información extractada.
Este articulo presenta una clasificación de estructuras de tablas HTML, la descripción de las formas de organización y presentación de información dentro de tablas, así mismo, un conjunto de reglas y heurísticas para la extracción de información contenida en tablas HTML, el diseño de la DTD intermedia que facilitara la generación de esquemas XML. Aquí se adopta XML como formato de salida del proceso de extracción, por presentar una alta adaptabilidad, accesibilidad e interoperalidad en diferentes entornos informáticos, y porque la W3C lo ha constituido en un estándar para el desarrollo de aplicaciones Web.

2. ESTRUCTURAS DE TABLAS

Para esta investigación una estructura de tabla corresponde a la forma de distribución de los títulos de tablas y títulos de columnas con su numero de columnas y líneas. Las tablas con numero de líneas M > 1 y numero de títulos N>1, son generalmente tablas bien formadas.
Dentro de la investigación se ha logrado determinar tres diferentes tipos de estructura de tablas.
El primer tipo de tabla encontrado, se define como una tabla con estructura simple, esta estructura de tabla se caracteriza por presentar las siguientes etiquetas:

- Las tablas son definidas por las etiquetas <table> ... </table>.
- Las etiquetas <TR> y </TR>, definen cada una de las líneas de la tabla.
- Las etiquetas <TD> y </TD>, definen cada uno de los elementos contenidos en una celda.
- La etiqueta <TBODY> permite agrupar un conjunto de elemento para la tabla.

Titulo de tabla				
1er titulo	**2do titulo**	**N esimo titulo**	**…**	**…**
.			…	…
.				

Figura 1. Estructura de una tabla simple

Un segundo tipo de estructura de tabla se ha denominado estructura de tablas con reagrupamientos, este tipo de tabla se caracteriza por presencia de los atributos de agrupamientos de filas y columnas: Rowspan y Colspan en las etiquetas **<tr> <td></td></tr>, o <tr> </tr> <th></th></tr>.**
En la *figura 2.* se presenta una estructura de tabla con reagrupamientos en filas y columnas. Este tipo de estructura es complejo y requiere para su interpretación y análisis, una conversión a formato de tabla simple (ver *figura 1)*

Titulo de Columna 1		Titulo de Titulo de Columna **2**	
		Columna **colspan 1**	Columna **colspan 2**
Titulo **Rowspan**	**Rowspan 1**	…	…
	Rowspan 2	...	...

Figura 2. Estructura de una tabla compleja

Las estructuras de tabla que presenten reagrupamientos son clasificadas en tres sub-grupos, y están definidas por las siguientes reglas lógicas:

- Solamente la presencia de: {<TD Colspan=n>} v {<TH Colspan=n>}, para una tabla que presente columna con reagrupamientos el valor mínimo por defecto será n=1.
- Solamente la presencia de {<TD Rowspan=m>} v {<TH Rowspan=m>}, el

valor mínimo por defecto en una tabla que presente una sola fila será m=1.

- La presencia de {<TD Colspan=n> v <TH Colspan=n>} & {<TD Rowspan=m> v <TH Rowspan=m>}

3. TIPOS DE DISTRIBUCION Y REPRESENTACION DE INFORMACION

Una vez determinados los tipos de estructura de tablas, el paso siguiente es el análisis de la distribución de información dentro de una tabla HTML.
El primer tipo de distribución de información dentro de una tabla, se caracteriza por el modelo tradicional de aridad.
Aquí una tabla esta definida por el contenedor <table>...</table>, las líneas están definidas por el contenedor <tr>...</tr>, y a su vez cada línea esta constituida por celdas individuales definidas por el contenedor <td>...</td> v <th>...</th>.

Caloric Value of Alcoholic Beverages

Alcoholic Beverage	Amount	Calories
Beer	12 oz (approx. 330 ml)	178
Wine	60 ml	63
Port Wine	60 ml	95
Gin (1 peg)	30 ml	73
Rum (1 peg)	30 ml	73
Whisky (1 peg)	30 ml	73
Vodka (1 peg)	30 ml	73
Brandy	30 ml	77

Figura 3. Distribución de información de forma clásica

En (F*igura 3.*) El contenido de cada etiqueta <td>...</td> v <th>...</th>, son caracteres de tipo alfa-numérico, con la posibilidad de encontrar caracteres especiales. La presencia de este tipo de distribución de información puede estar presente con una estructura simple (*figura 1)*), o bien en una estructura de tabla compleja con reagrupamiento de filas y columnas (*figura 2.*).
Nótese que cada celda de cada columna de la tabla tiene una relación uno a uno con las demás celdas de su respectiva fila.

Un segundo tipo de distribución establece una tabulación entre elementos de columnas, ha esta distribución se le ha denominado decallage[1], aquí la característica principal es la presencia del atributo .
Una etiqueta td>...</td> v <th>...</th> con tabulado entre sus datos esta representado por la siguiente figura:

Item	Approximate pH
Anchovies	6.50
Anchovies, stuffed w/cappers, in olive oil	5.58
Apple, eating	3.30 - 4.00
Apples	
Delicious	3.90
Golden Delicious	3.60
Jonathan	3.33
McIntosh	3.34
Winesap	3.47
Juice	3.40 - 4.00
Sauce	3.30 - 3.60
Apple, baked with sugar	3.20 - 3.55

Figura 4. Distribución de información con tabulado.

Un tabulado es una forma de jerarquías de datos, él permite el agrupamiento de sub-grupos o sub-clases.
Un tercer tipo de distribución de información se caracteriza por la presencia de contenedores (parágrafos) en las etiquetas <td></td> v <th></th>, un contenedor puede soportar grandes cadenas de caracteres, este tipo de distribución puede considerarse no estructurada.

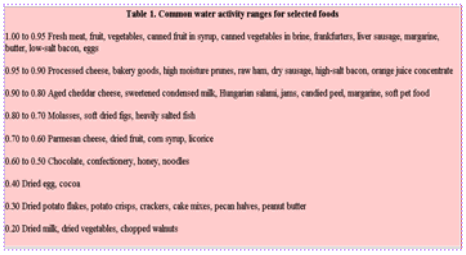


Table 1. Common water activity ranges for selected foods

1.00 to 0.95 Fresh meat, fruit, vegetables, canned fruit in syrup, canned vegetables in brine, frankfurters, liver sausage, margarine, butter, low-salt bacon, eggs

0.95 to 0.90 Processed cheese, bakery goods, high moisture prunes, raw ham, dry sausage, high-salt bacon, orange juice concentrate

0.90 to 0.80 Aged cheddar cheese, sweetened condensed milk, Hungarian salami, jams, candied peel, margarine, soft pet food

0.80 to 0.70 Molasses, soft dried figs, heavily salted fish

0.70 to 0.60 Parmesan cheese, dried fruit, corn syrup, licorice

0.60 to 0.50 Chocolate, confectionery, honey, noodles

0.40 Dried egg, cocoa

0.30 Dried potato flakes, potato crisps, crackers, cake mixes, pecan halves, peanut butter

0.20 Dried milk, dried vegetables, chopped walnuts

Figura 5. Distribución de información con Parágrafos.

La (*Figura 5*) es un ejemplo clásico de este tipo de distribución de información, aquí la tabla presenta solamente una columna, es decir un <tr></tr> que contiene la etiqueta <td></td>, mientras que el contenido de la etiqueta <td></td>, estará constituido por contenedores { } v { ... }, que define cada una de las líneas de la tabla.

4. LA DTD Y EL DOCUMENTO XML INTERMEDIARIO

La DTD o (Definición de Tipo de Documento), determina las reglas que debe cumplir cada uno de los elementos y atributos del documento XML[2], fue creada para obtener el XML intermediario y agrupa las siguientes características:

- Cuales son los elementos permitidos en el documento XML.
- El contenido que puede tener los elementos.
- Que atributos pueden estar asociados a cuales elementos.
- Cuales son los valores permitidos por los atributos.

De acuerdo a las características anteriores, se ha determinado los siguientes elementos para la DTD:

- Se ha definido un elemento global denominado tabla <!ELEMENT tableau (titre, Num-col, contenu)>, que contiene tres elementos: titulo, Numero de columnas y contenido.
- El elemento titulo dado por <!ELEMENT titre (titre-tableau?, titre-col+)>, contiene los sub-elementos: Titulo de tabla, titulo de columna.
- El elemento Numero de columnas <!ELEMENT Num-col ANY>, determina el numero de columnas que presenta la tabla HTML.
- El elemento Contenido <ELEMENT contenu (linge+)>, representa el cuerpo de la distribución de información de la tabla; este elemento determina cada una de las filas (<tr>…</tr>) contenidas en una tabla.
- El atributo (linge+)>, presenta el atributo (case+)> que determina cada uno de los datos contenidos en una línea (<td>)

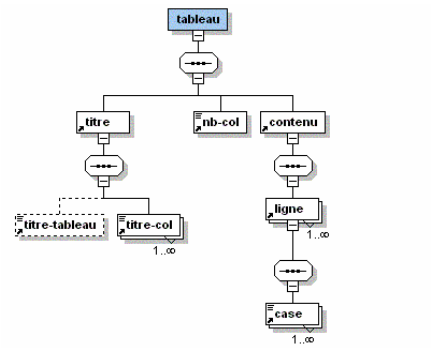


Figura 6. XML Schema de la DTD Generada.

La DTD generada a partir del esquema mostrado en la *Figura 6.* es la siguiente:

```
<?xml version="1.0" encoding="UTF-8"?>
<!--DTD generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)-->
<!--Comment describing your root element-->
<!ELEMENT tableau (titre, nb-col, contenu)>
<!ELEMENT titre (titre-tableau?, titre-col+)>
<!ELEMENT nb-col ANY>
<!ELEMENT contenu (ligne+)>
<!ELEMENT titre-tableau ANY>
<!ELEMENT titre-col ANY>
<!ELEMENT ligne (case+)>
<!ELEMENT case ANY>
```

En la DTD se observa que algunos de sus elementos presentan operadores, estos operadores (Any, ?,+) permiten establecer reglas de restricción para estos elementos, por ejemplo el elemento <!ELEMENT nb-col ANY>, que determina el numero de columnas de una tabla, señala que el puede contener todo tipo de datos. En tanto el operador (+) contenido en el elemento titre-col+, determina que esa información debe estar presente al menos una vez. El elemento titre-tableau? Contiene el operador (?) Y determina que la presencia de este elemento es opcional.

6. EL PROCESO DE EXTRACCION DE INFORMACION

Una vez generada la DTD intermediaria, el paso siguiente es el establecer el conjunto de reglas y heurísticas para la extracción de la información a partir del conjunto de tablas que pueden pertenecer o no a un determinado dominio.
La entrada al algoritmo es un archivo con el código fuente HTML, este archivo puede contener o no una o mas tablas, la salida del algoritmo corresponde a un archivo de tipo XML, que contiene la información extractada de una o varias tablas HTML.
El algoritmo de extracción reconoce e interpreta los siguientes pasos:

- La detección y selección de una o mas tablas bien formadas.
- La detección ,selección y extracción de títulos de tablas
- La detección del número de columnas.
- La detección y extracción de títulos de columna de cada tabla.
- La extracción de información contenida en las filas y celdas
- La generación del documento XML, de acuerdo a las características propias de cada tabla.

Para el algoritmo de extracción, una tabla es valida si esta contenida dentro del dominio de tablas descritas en el capitulo numero 3, las reglas para la selección y restricción de una o mas tablas HTML esta dada por el siguiente conjunto conflicto[3]:
R1 Si archivo HTML contiene las etiquetas <tablea>...</table>, entonces asignar tabla al conjunto {tabla i}

R2 Si {tabla i} $\subset$ una columna n=1 y m numero filas y todas sus filas contienen referencias “HREF”, $\Rightarrow$ “tabla no valida”.

R3 Si {tabla i} ⊂ si y solo si una columna y una fila ⇛ “tabla no valida”.

R4 Si {tabla i} ⊂ Colspan=n y ∃ filas <tr>...</tr> de la tabla con numero de celdas <td>...</td> menor o mayor a Colspan n. ⇛ “tabla no valida”.

R5 Si {tabla i} ⊂ doble Rowspan para una misma fila <tr>...</tr>. ⇛ “tabla no valida”.
Las tablas que no cumplan con estas restricciones serán consideradas como tabla valida para extractar información.

La detección de un titulo de tabla es un proceso un tanto mas complejo, requiere de un análisis un tanto mas detallado, la presencia de un titulo de una tabla no siempre es transparente, algunas tablas lo presentan directamente, pero en otras tablas es necesario detectarlo y seleccionarlo.
Aquí se presenta las reglas para la detección y selección del titulo de una tabla:

R1 Si ∀ {tabla $_i$}: ∃ <CAPTION> “X ”</CAPTION> ⇛ x= <titulo de tabla>
sino

R2 Si {tabla $_i$} ⊂ {<TD Colspan=n “X”> ∨ <TH Colspan=n “X”>} ⇛“X” =<titulo de tabla>
sino

R3 Si <tr>...</tr> anterior a <table> ... </table> ⊂ <h1> ∨ <h1> ∨ <h2> ∨ <h3> ∨<h4> ∨<h5> ∨ <h6> ⇛<h1>...<h6>=<titulo de tabla>

R6 Si ∀ {tabla $_i$}: ∃ Num-Columnas = 1 ∧ {<tr$_1$><td> ”X” </td></tr$_1$>}⇛”X” =<titulo de tabla>
sino

R5 Si ∃ titulo de pagina</TITLE>“X”</TITLE> ⇛ “X” = <titulo de tabla>
sino

R6 = <titulo de tabla> = “tabla sin titulo”

La detección y selección de los títulos de columnas de tabla presenta un tratamiento similar al anterior, aquí también se asume que la tabla seleccionada {table i} fue filtrada y por ende esta bien formada para iniciar su análisis de extracción. Se han definido las siguientes reglas generales para determinar, seleccionar y extraer la información de las columnas de una tabla.

R1 Si ∀{tabla $_i$}: ∃{<TD Colspan=n> ∨ <TH Colspan=n>} ⇛{<tr$_1$><td$_i$>“X$_i$”</td$_i$></tr$_1$>}= <titulo de columna n> ∨ {<tr$_1$><th$_i$>“X$_i$”</th$_i$></tr$_1$>} = <titulo de columna n>
sino

R2 Si {tabla $_i$} ⊂ <thead> ⇛{<tr$_1$><th$_i$“X”</th$_i$ ></tr$_1$>} = <titulo de columna n>
sino

R3 Si {tabla $_i$} ⊄ <thead> ⇛{<tr$_1$><td$_i$“X”</td$_i$ ></tr$_1$>} = <titulo de columna n>

R4 Si ∀ {tabla $_i$}: ∃ Num-Columnas = 1 ∧ {<tr$_1$><td> ”X” </td></tr$_1$>}⇛”X” =<titulo de columna>

El numero de filas de una tabla se determina de acuerdo al numero de contenedores o etiquetas <tr>...</tr>, mientras que el numero de columnas se obtiene a traves del numero de etiquetas <td>...</td> contenidas en la primer linea <tr>...</tr> de la tabla.

7. RESULTADOS

La presencia de tablas con estructuras complejas es común en los documentos html, para su análisis es necesario efectuar un mapeo a través de jerarquías semánticas. Los elementos colspan y rowspan juegan un papel importante en este tipo de conversión de estructura de tabla, este mapeo semántico se realiza para obtener una estructura de tabla mas simple, pero semánticamente igual, una vez obtenida la tabla simple a partir de la estructura compleja, se procede al análisis de la tabla, para ejecutar el proceso de extracción antes descrito.

Population		Moyenne	
		Hauteur en cm	Espérance de vie
Sexe	Masculin	1,78	70
	Féminin	1,70	80

Population.Sexe	Moyenne.Hauteur en cm	Moyenne.Esperance de vie
Masculin	1,78	70
Feminin	1,70	80

Figura 7. Generación de una tabla con estructura simple, a partir de una estructura compleja.

En la *figura 7,* se observa la simplificación de una tabla compleja a una estructura simple.
El proceso de extracción de información a partir del contenido de tabla, identifica los tres diferentes tipos de distribución descritos en el capitulo 3, la estructura XML es flexible y permanece constante para los tres tipos de distribución de información. En el caso de una distribución de información con tabulados la salida XML permite determinar las instancias entre los elementos agrupados por el tabulado,

en la *figura 4,* se observa que la clase Apples involucra las subclases delicious, golden delicius, hasta la subclase souce, por tanto podemos definir que delicious, golden delicius etc; son tipos de Apple. De esta manera la equivalencia semántica para este tipo de tabla es de la forma: Apples.delicious, Apples. golden delicius etc...
Un ejemplo de salida XML para la tabla de la *figura 4* es:

```
<tableau>
        <titre>
                <titre-tableau>FDA/CFSAN: Approximate pH of
Foods and Food Products</titre-tableau>
                <titre-col>Item</titre-col>
                <titre-col>Approximate pH</titre-col>
        </titre>
<nb-col>2</nb-col>
<contenu>
                <ligne>
                        <case>Anchovies </case>
                        <case>6.50</case>
                </ligne>
                <ligne>
                        <case>Anchovies, stuffed
                        w/cappers, in olive oil </case>
                        <case>5.58 </case>
                </ligne>
                <ligne>
                        <case>Apple, eating</case>
                        <case>3.30 - 4.00</case>
                </ligne>
                <ligne>
                    <case>Apples .Delicious</case>
                      <case>3.90 </case>
                </ligne>
                <ligne>
                        <case>Apples .Golden Delicious
                        </case>
                        <case>3.60 </case>
                </ligne>
                </ligne>
                …….
        </contenu>
</tableau>
```

Nótese que en el documento XML se ha eliminado fila Apple de la tabla html, y solo se presenta la clase y sus sub clases para este grupo de elementos tabulados. En tanto para una tabla que presente parágrafos <p> como forma de distribución de información, estos serán considerados en el proceso de extracción como una fila de la tabla a pesar que estén contenidos dentro de una misma celda <td></td>. La información extractada a partir de la tabla de estructura compleja visualizada en la *figura 7* , se muestra a continuación:

```
<tableau>
        <titre>
                <titre-tableau>Don't have title</titre-tableau>
                <titre-col>Population.Sexe</titre-col>
                <titre-col>Moyenne.Hauteur en cm </titre-col>
                <titre-col>Moyenne.Esperance de vie</titre-col>
        </titre>
<nb-col>3</nb-col>
<contenu>
                <ligne>
                        <case>Masculin</case>
                        <case>1,78</case>
                        <case>70</case>
                </ligne>
                <ligne>
                        <case>Feminin</case>
                        <case>1,70</case>
                        <case>80</case>
                </ligne>
        </contenu>
</tableau>
```

5. CONCLUSIONES

Este articulo presento una aproximación hacia la extracción e integración de información contenida en tablas de tipo HTML.
Se ha elaborado un estudio y posterior clasificación de tablas según su estructura y complejidad que han permitido establecer y definir una serie de reglas heurísticas para lograr extraer la información contenida en las tablas de este tipo.
Para tal efecto se ha implementado un prototipo informático que permite capturar los documentos WEB, detectar la presencia de tablas y filtrar las tablas que sean validas para el proceso, además permite visualizar la salida de información a través de documentos tipo XML.
Después de evaluar los resultados obtenidos a través del prototipo de extracción, podemos afirmar que los resultados son satisfactorios y de alta calidad, eso permite mirar cuales serán las tareas futuras en la investigación.
El siguiente paso será el enriquecimiento semántico de los documentos XML generados. Este proceso se realizara a través del uso de una Ontología en un dominio por determinar.

6. REFERENCIAS BIBLIOGRAFICAS

[1.] J.C Stuhtmann " Grand Livre HTML" .Ed Micro Application 1999
[2.] H. Maruyama "XML y Java" . Ed Prentice Hall 1999
[3.] G.F. Luger "Artificial Intelligence. Structures and Strategies for Complex Problem Solving". Benjamin/Cummings Publishing, 1992.

7. AUTOR

Marco-Javier Suárez B.
Ingeniero de Sistemas Uniboyaca 1997, Esp. Sistemas Informáticos Universidad Nacional 2001, Actualmente Candidato a Doctor en Informatica Université Paris XI.

Profesor auxiliar
Facultad de Ingeniería
UPTC, Residences Fleming Paris-France,
Telefono: ++ 33 1 69 15 34 76.
suarez@lri.fr, http://www.lri.fr/~suarez