Modelo de Extracción Automática de Información a partir de Tablas HTML

Marco Javier Suárez Barón

Candidato a Doctor en Informática

Laboratoire de Recherche en Informatique LRI
Université Paris XI sud
Orsay 9500-Francia
suarez@lri.fr
javiers74@hotmail.com

Dirigido a: Workshop de Agentes y Sistemas Inteligentes

Resumen. Este trabajo presenta un modelo de extracción e integración de información contenida dentro de tablas de tipo HTML, el modelo de extracción de información se apoya en un conjunto de heurísticas y reglas de deducción. Para determinar este conjunto de reglas se determino una detección rigurosa de tablas y elementos de tablas HTML según su tipo de estructura y complejidad. El propósito de la investigación es extractar e integrar la información con el propósito de generar información estructurada bajo la forma documentos de tipo XML.

Palabras Clave: Extracción de Información, Wrapper, Heurísticas, DTD, XML, HTML

1 Introducción

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 Metodología

El proceso abarca un estudio minucioso de los diferentes tipos de estructuras y elementos de tablas HTML la detección de las formas distribución de información dentro de la tabla, a partir de este análisis se realiza una clasificación de estructuras de tablas tendiente a determinar el conjunto de heurísticas y reglas de deducción que permitirán obtener la información deseada y su futura integración.

2.1 Detección del conjunto de estructuras y elementos de tablas HTML

El primer tipo de tabla detectado, se define como una tabla con estructura simple, ver Fig. 1, esta estructura de tabla se caracteriza por presentar las siguientes etiquetas:

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

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

Fig. 1. Distribución de información de forma clásica

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 la Fig.1 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.

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

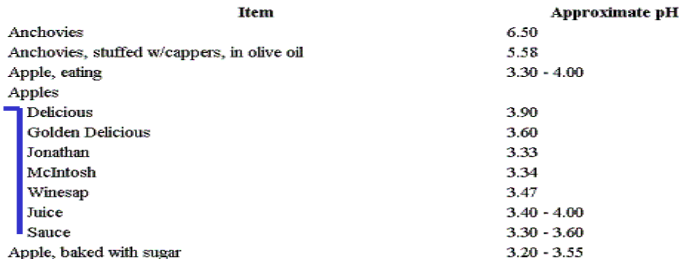
Fig. 2. Estructura de una tabla compleja

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

1. 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.

2. 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.
3. La presencia de {<TD Colspan=n> v <TH Colspan=n>} & {<TD Rowspan=m> v <TH Rowspan=m>}

En cuanto a tipo de distribución; existe una tabulación entre elementos de columnas, ha esta distribución se le ha denominado decallage[1], aquí la característica principal es la presencia del atributo , Un tabulado es una forma de jerarquías de datos, él permite el agrupamiento de sub.-grupos o sub.-clases. Ver Fig. 3.



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

Fig. 3. Distribución de información con tabulado.

El siguiente tipo de distribución de información se caracteriza por la presencia de contenedores (parágrafos) en las etiquetas <td></td> v <th></th>, ver Fig. 4. Un contenedor puede soportar grandes cadenas de caracteres, este tipo de distribución puede considerarse no estructurada. Un ejemplo clásico de este tipo de distribución de información son las tablas que presentan 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.

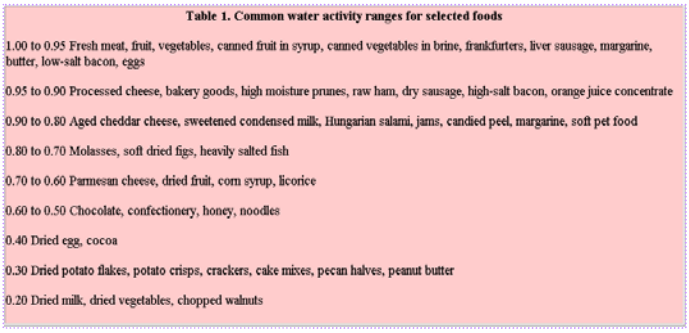


Table 1. Common water activity ranges for selected foods

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

Fig. 4. Distribución de información con Parágrafos.

2.2 Diseño de la DTD y documentos XML

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:

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

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,a su vez el elemento Número 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>).

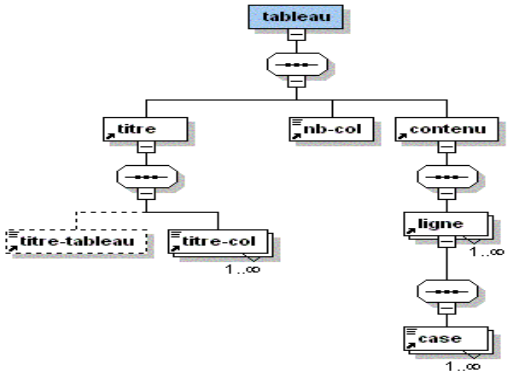


Fig. 5. XML Schema de la DTD Generada.

La DTD generada a partir del esquema mostrado en Fig. 5 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.

2.3 El proceso de extracción de información

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 wrapper, que contiene la información extractada de una o varias tablas HTML.

El algoritmo de extracción reconoce e interpreta los siguientes pasos:

1. La detección y selección de una o mas tablas bien formadas.
2. La detección , selección y extracción de títulos de tablas
3. La detección del número de columnas.
4. La detección y extracción de títulos de columna de cada tabla.
5. La extracción de información contenida en las filas y celdas
6. La Integración de tablas y 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 la sección 3.1, el algoritmo permite determinar si el documento HTML de entrada presenta tablas o no, si este archivo fuente presenta una o mas posibles tablas; estas son almacenadas en un conjunto de tablas denominadas {tablas i}. Las reglas Ri para la selección y restricción de una o mas tablas HTML esta dada por el siguiente conjunto conflicto[4]:

R1 Si archivo HTML $\subset$ las etiquetas {<table>...</table>} $\rightarrow$ asignar tabla al conjunto {tabla i}
R2 Si {tabla i} $\subset$ columna n=1 $\subset$ m numero filas $\wedge$ $\forall$ fila(i) : $\exists$ referencias(href.)$\rightarrow$ "tabla no valida"
R3 Si {tabla i} $\subset$ si y solo si columnas n=1 $\wedge$ filas m=1$\rightarrow$ "tabla no valida"
R4 Si {tabla i} $\subset$ Colspan = n $\wedge$ $\forall${tabla i} : $\exists$ filas(<tr>...</tr>) con numero de celdas (<td>...</td>) < o> a Colspan = n. $\rightarrow$ "tabla no valida"
R5 Si {tabla i} $\subset$ doble Rowspan para una misma fila (<tr>...</tr>) $\rightarrow$ "tabla no valida"

La detección de un titulo de tabla requiere un análisis 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. A continuación se presenta las reglas de deducción para la detección y selección del titulo de una tabla:

R1 Si $\forall$ {tabla i}: $\exists$ <CAPTION> "X "</CAPTION>$\rightarrow$ "X"= <titulo de tabla> si-no
R2 Si {tabla i} $\subset$ {<td Colspan=n "X"> v <th Colspan=n "X">} $\rightarrow$"X" =<titulo de tabla> si-no
R3 Si $\exists$ fila(<tr>...</tr>) anterior a <table> ... </table> $\wedge$ fila(<tr>...</tr>) $\subset$ {(<h1>"X"</h1>) $\vee$ (<h2>"X"</h2>) $\vee$ (<h3>"X"</h3>) $\vee$ (<h4>"X"</h4>) $\vee$ (<h5>"X"</h5>) $\vee$ (<h6> "X"</h6>)$\rightarrow$ "X"=<titulo de tabla>
R6 Si $\forall${tabla i}: $\exists$ Num-Columnas = 1 $\wedge$ {tabla i} $\subset$ fila(<tr1><td> "X" </td></tr1>)$\rightarrow$"X" =<titulo de tabla> si-no
R5 Si $\exists$ etiqueta(</TITLE>"X"</TITLE>) $\rightarrow$ "X" = <titulo de tabla>sino
R6 = <titulo de tabla> = "tabla sin titulo"

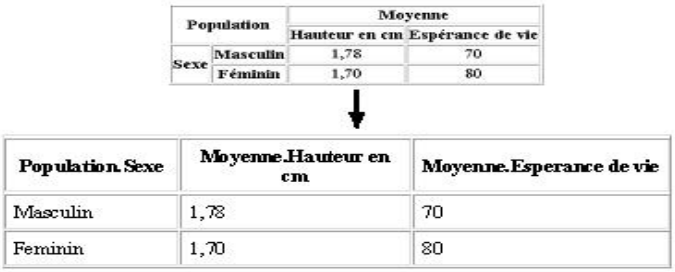
El paso siguiente es la detección y selección de los títulos de columnas de tabla, aquí también se asume que la tabla seleccionada {tabla i} fue filtrada y por ende esta bien formada para iniciar su análisis de extracción.

R1 Si $\forall${tabla i}: $\exists${<td Colspan=n> $\vee$ <th Colspan=n>} $\wedge$ $\exists${<tr1><tdi>"Xi"</tdi></tr1>}$\rightarrow$"Xi"= <titulo de columna n> $\vee$ $\exists${<tr1><thi>"Xi"</thi></tr1>} $\rightarrow$ "Xi"= <titulo de columna n> sino
R2 Si {tabla i} $\subset$ <thead> $\wedge$ $\exists${<tr1><thi"X"</thi ></tr 1>}$\rightarrow$"X" = <titulo de columna n> si-no
R3 Si {tabla i} $\not\subset$ <thead> $\wedge$ $\exists$ {<tr1><tdi"X"</tdi ></tr 1>} $\rightarrow$ "X" <titulo de columna n>

R4 Si ∀ {tabla i}: ∃ Num-Columnas = 1 ∧ ∃{<tr1><td> ”X” </td></tr1>}→”X” =<titulo de columna>

3 Resultados Experimentales

La presencia de tablas con estructuras complejas es común en diferentes 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[3] 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 i}, 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

Fig. 6. Generación de una tabla con estructura simple, a partir de una estructura Compleja.

En la Fig. 6, 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 la sección 2.1, la estructura XML es flexible y permanece constante para toda estructura de tabla y tipo de distribución de información. La información extractada e integrada de la la Fig. 3 se visualiza en la Tabla 1.

```
<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>
```

Tabla 1. Wrapper XML Generado a partir de la tabla HTML de la Fig. 4.

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 Fig. 2 , 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 Fig. 2 se visualiza en la Tabla 2.:

```
<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>
```

Tabla 2. Wrapper XML Generado a partir de la tabla HTML de la Fig. 2.

3 Conclusiones y Trabajos Futuros de Investigación

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 de sistema inteligente a nivel wrapper que permite capturar los documentos WEB, detectar la presencia de tablas y filtrar las tablas que sean validas para el proceso, además permite extraer, generar y visualizar automáticamente 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 proyectar que las tareas futuras en la investigación se dirigirá hacia el enriquecimiento semántico de los documentos XML generados. Este proceso se realizara a través del uso de una Ontología en un dominio por determinar.

Referencias

[1] J. Hammer y H. Garcia Molina. Extracting Semistructured Information from the Web. *In Proceding of the Workshop on Management of Semistructured data*, May 2002.

[2] M. Fernández y D. Suciu. SilkRoute: Trading between relations and XML. *In Procedings of 9th Intl. Conf. On Word Wide Web*, 2000

[3] S. Bergamaschi y S. Castano, Semantic Integration of Semistructured and structured data sources. SIGMOD record, pp 54-59, 2000.

[4] G.F. Luger , *Artificial Intelligence. Structures and Strategies for Complex Problem Solving*. Benjamin/Cummings Publishing, 1992.