ONTOLOGÍAS PARA SISTEMAS MEDIADORES

GLORIA LUCÍA GIRALDO GÓMEZ[1]

Resumen

En el contexto de la informática, los sistemas mediadores son definidos como «unos programas de computador (software) que explotan el conocimiento relativo a un conjunto de datos para crear información destinada a las aplicaciones de los usuarios» (Wiederhold, 1992). El objetivo de los mediadores es simplificar, abstraer, reducir, integrar y combinar datos que se encuentran en fuentes diferentes, para responder mejor a las consultas de los usuarios. Un sistema mediador es específico a un dominio de aplicación, éste es descrito por un esquema llamado esquema global u ontología del dominio. El rol de la ontología es central, ya que es el soporte para la expresión de las consultas (queries) de los usuarios. Igualmente, es a través de este esquema que las distintas fuentes de datos se interconectan. La construcción automatizada de los sistemas mediadores puede ser abordada, en un principio, automatizando la construcción de una parte fundamental de estos sistemas, como es la ontología. La proposición de métodos y técnicas para construir, lo más automáticamente posible, una ontología de dominio para un sistema mediador, es el tema de la investigación que se llevó a cabo en el marco de una tesis doctoral[2], en el LRI (Laboratoire de Recherche en Informatique) de la

1 Ingeniera de Sistemas, Universidad de Antioquia. Especialista en Ciencias Electrónicas e Informática con énfasis en Bases de Datos, Universidad de Antioquia. Magíster en Teoría e Ingeniería de Bases de Datos, Universidad de Paris I Panteón-Sorbona (Francia). PhD en Informática de la Universidad de Paris XI Orsay (Francia). Asesora de líneas de investigación aplicada del ITM. Coordinadora del grupo de investigación GIT del ITM, dedicado a la integración de soluciones con tecnología de información y comunicación.
E-mail: gloriagiraldo@itm.edu.co.

2 Construction automatisée de l'ontologie de systèmes médiateurs: application à des systèmes intégrant des services standards accessibles via le Web, Gloria Lucía Giraldo Gómez, Université Paris Sud XI, 2005.

Universidad Paris Sur XI (Orsay-Francia). La solución propuesta está compuesta de tres etapas: extracción de los elementos que componen la ontología, a partir de un conjunto de DTDs (Document Type Definition), la estructuración de los elementos extraídos y la representación de estos elementos. Las técnicas propuestas han sido implementadas en un sistema llamado OntoMedia. El objetivo de este artículo es mostrar los primeros resultados de dicha investigación. Debido a limitaciones de espacio, se presentarán la primera y segunda etapa. La tercera será el objeto de una próxima publicación.

Dado que las ontologías tienen actualmente gran aplicabilidad, dejan muchos temas abiertos para la investigación. Como una consecuencia de este hecho estamos desarrollando una sublínea adscrita al grupo de investigación en integración de soluciones con tecnologías de información y comunicación –GIT del INSTITUTO TECNOLÓGICO METROPOLITANO. Los aspectos importantes que justifican el continuar investigando en este campo son introducidos en la parte final de este artículo.

Palabras clave

Ontologías, sistemas mediadores, integración de información, Web Semántica.

Abstract

In Computer Sciences, mediators systems are defined as «computer software that exploit the knowledge related to a set of data in order to create information for users' applications» (Wiederhold, 1992).

The objective of the mediators is to simplify, to abstract, to reduce, to integrate and to combine data from different sources in order to give better answers to user's queries. A mediator system is specific to an application domain. This is described by a global schema or domain ontology. The role of ontology is central to the schema because it is the support for the expression of users' queries. Also, it is through this schema that the different sources of data are interconnected.

The automated construction of mediators systems can be initially approached by automating the construction of a fundamental part of these systems, as ontology is. Some methods and techniques to construct, as automatically as possible, a domain ontology for a mediator system, were the topic of investigation carried out for a PhD thesis at LRI (Laboratoire de Recherche en Informatique) in the Paris Sud XI University (Orsay-France).

The proposed solution is made up of three stages: The first one, the extraction of the elements that make up the ontology, starting with a set of DTDs (Document Type Definition). Second, the organisation of these elements and the last stage is their representation. The techniques that are suggested have been implemented in a system named OntoMedia. The objective of this article is to show the first results of such research. Due to space limitations, the first and the second stage are presented. The third stage will appear in a publication to follow.

Considering that ontologies have a great deal of applications at the moment, a lot of subjects are left for future research. As a consequence, we are developing a sub-line, ascribed to the team that is doing research in integration of solutions with informatics and communication technologies - a research group at ITM. The important aspects that justify keeping on research in this field are mentioned in the final part of this article.

Key words

Ontology, Mediators systems, Information integration, Semantic Web.

1. Introducción

El contexto de utilización de los sistemas mediadores es la integración de información. Existen dos grandes acercamientos a la integración de información: el almacén de datos (Data Warehouse) (Chaudhuri et Dayal, 1997) y el mediador (Wiederhold, 1992). En el primero, la integración consiste en construir grandes bases de datos reales que reagrupan la información necesaria para las aplicaciones. En los mediadores la integración se basa en la explotación de vistas abstractas que describen el contenido de las diferentes fuentes de información. La presente investigación se sitúa en el contexto de éste último acercamiento.

Dos arquitecturas de sistemas mediadores existen hoy: centralizados y descentralizados. Los primeros sistemas que se han implementado son de arquitectura centralizada. La arquitectura descentralizada se basa en el paradigma par-a-par (peer-to-peer o P2P). El interés de este trabajo se centra en los sistemas mediadores de arquitectura centralizada y, particularmente, en el sistema mediador PICSEL[3] (Producción de Interfaces con base en Conocimiento para Servicios en Línea). PICSEL es igualmente el nombre del proyecto que se ha desarrollado en colaboración con la empresa «France Telecom R&D». Su desarrollo fue en varias fases. En la primera (1997-2000), llamada PICSEL1, el objetivo ha sido proveer un ambiente declarativo de desarrollo de sistemas mediadores centralizados. En la segunda (2000-2003), llamada PICSEL2, el objetivo ha consistido en proponer soluciones innovadoras haciendo que PICSEL1 dé el salto a la Web y permitir la integración de un número importante de fuentes múltiples y heterogéneas. En esta segunda fase se desarrolló la investigación que se describe en el presente artículo. Actualmente, algunos investigadores del laboratorio de investigación en informática - LRI (http://www.lri.fr), en Francia, desarrollan PICSEL3, el cual corresponde la tercera fase del proyecto.

3 Production d'Interfaces à base de Connaissances pour des Services En Ligne.

El aporte de los sistemas mediadores cubre diferentes aspectos: en primer lugar, permite descubrir las fuentes que son pertinentes para responder las consultas de los usuarios. Además ayuda a acceder a dichas fuentes, que normalmente son heterogéneas, lo que evita que el usuario tenga que interrogarlas una a una utilizando el vocabulario y el lenguaje propios de cada una de ellas. Finalmente, los sistemas mediadores combinan automáticamente las respuestas parciales obtenidas de varias fuentes, para entregar una respuesta global al usuario.

En los sistemas mediadores, el papel de la ontología es central, pues ella modela el dominio de aplicación del sistema y brinda un vocabulario estructurado que sirve de soporte para la expresión de las consultas de los usuarios. Además, ella establece, indirectamente, una relación entre las fuentes, puesto que el contenido de cada fuente es descrito con la ayuda de un mismo vocabulario: el de la ontología.

El término Ontología (con O mayúscula) corresponde en sus orígenes, a una disciplina filosófica: «parte de la metafísica que trata del ser en general y de sus propiedades trascendentales» (http://www.rae.es). Etimológicamente viene de los términos griegos «ontos» y «logos» (http://www.webdianoia.com/glosario), que quieren decir el tratado del ser.

El concepto de ontología (con o minúscula) ha sido introducido en la informática por la comunidad «inteligencia artificial» a principios de los años 90 (Neches at al., 1991). Este concepto fue retomado luego en los trabajos de Ingeniería del Conocimiento (IC) y fue redefinido en función de su utilización, dando lugar a la fusión de varias interpretaciones. Nicolás Guarino (1995) da 7 interpretaciones diferentes del concepto de ontología.

La definición de ontología más referenciada en la literatura ha sido la de Gruber (1993): «una ontología es una especificación explícita de una conceptualización». En 1997, esta definición fue modificada por Borst de la manera siguiente: «una ontología es una especificación formal de una conceptualización compartida» (Borst, 1997). Fundiendo estas dos definiciones se tiene que «una ontología

es una especificación explícita y formal de una conceptualización compartida». Una conceptualización es el conjunto de entidades pertinentes de un dominio de estudio, de sus propiedades y de las relaciones que existen entre ellas. Es una vista abstracta y simplificada del mundo que se desea representar. Esta especificación es *explícita*, porque los tipos de los conceptos utilizados, las relaciones entre ellos, las restricciones que ellos deben satisfacer y sus propiedades son explícitamente definidas. Ella se dice *formal* porque debe ser el objeto de tratamientos automáticos, y es *compartida* porque refleja el consenso de una comunidad.

Actualmente, la utilización de este concepto es muy amplia. Cada comunidad adopta su propia interpretación, pero como regla general todos coinciden en decir que se trata de la representación de una estructuración de un dominio, dominio en el sentido de área de conocimiento o área de interés. Las ontologías son utilizadas para catalogar y definir los tipos de las cosas que existen en un cierto dominio, así como sus relaciones y propiedades. Por ejemplo, una ontología del mundo empresarial usará conceptos como Venta, Compra, Transferencia, Pago, etc.; y relaciones como «una Transferencia corresponde a una Venta o a una Compra», «un Pago corresponde a una o varias Transferencias», etcétera.

Las formas de las ontologías son variadas. Algunas, por ejemplo, se limitan a ser simples jerarquías de conceptos. Es decir, un conjunto de conceptos (clases) ligados únicamente por relaciones de generalización/especialización o lo que es lo mismo por relaciones de subsunción (en inglés «is-a»). Supongamos, por ejemplo, las clases «Medio de transporte» y «Vehículo». La primera es una clase más general que la segunda, entonces la relación existente es: «Vehículo es un Medio de transporte».

En una ontología siempre habrá un conjunto de términos y una especificación de su significación. Estos términos unidos, bien sea, únicamente por relaciones de subsunción o por un conjunto de propiedades y de otras relaciones, como por ejemplo, relaciones propias al dominio.

Este artículo está estructurado de la siguiente manera: en la sección 2 se presenta el contexto y la problemática. En la sección 3, el método de construcción de la ontología. Ésta incluye la descripción de la metodología, una experimentación realizada y el análisis de los resultados. En la sección 4, las conclusiones. Finalmente, en la sección 5 se muestran, como perspectiva, algunos elementos de la formulación de una sublínea de investigación para ser desarrollada en el Instituto Tecnológico Metropolitano.

2. Contexto y problemática

Para explicar de manera intuitiva la utilidad de un sistema mediador, consideremos el siguiente ejemplo. Supongamos que un usuario busca en Roma un hotel cinco estrellas, para pasar algunos días en diciembre de 2006. Este usuario desea conocer la dirección y las características de los hoteles donde él se puede alojar. Simulemos la existencia de dos fuentes de información: F1 y F2. La primera es un archivo XML que contiene una lista de nombres de hoteles europeos con su respectiva dirección y una descripción de sus características. F2 es una base de datos relacional que contiene una lista de hoteles italianos, indicando la disponibilidad de cuartos por fecha. Un sistema mediador encuentra automáticamente que las fuentes F1 y F2 permiten responder a la consulta de este usuario. El mediador determina cuáles son los datos precisos que se deben extraer de cada fuente, de tal manera que se responda completamente a la consulta del usuario. La respuesta global y completa es construida por combinación de los datos obtenidos en cada fuente. Así, el usuario no tiene que consultar la fuente F2 para encontrar los hoteles italianos que tienen una habitación disponible en una fecha dada, y luego la fuente F1 para saber cuáles de esos hoteles son cinco estrellas y extraer su dirección.

En PICSEL1 el objetivo ha sido la construcción de un sistema mediador que permita «interrogar fuentes de información múltiples y heterogéneas relativas a un mismo dominio de aplicación». Un sistema mediador como PICSEL está compuesto de dos partes:

- Una parte genérica, es decir que puede ser utilizada en cualquier dominio de aplicación, ésta es el «Motor de consultas» (Goasdoue, 2001).
- Una parte específica al dominio, que corresponde a la(s) base(s) de conocimientos relativos al dominio del servidor. Una base de conocimientos se compone, a su vez, de una descripción del dominio del servidor, llamada también ontología del dominio y de las descripciones del contenido de las fuentes de datos por interrogar.

Una experiencia de construcción manual de una ontología para un sistema mediador fue realizada en la primera fase del proyecto PICSEL. En esta experiencia se concluyó que la creación de una ontología es un obstáculo grande para la construcción de los sistemas mediadores, aun más si se piensa desarrollar un sistema a gran escala. Se verificó, además, que el trabajo de modelación de una ontología es un trabajo largo y difícil. Fue por ello que en la segunda fase del proyecto PICSEL (PICSEL2), el objetivo perseguido ha sido el de proponer e implementar soluciones para automatizar, al máximo, la construcción de los sistemas mediadores, principalmente de la ontología, contribuyendo así a que estos sistemas den el tan ambicioso «salto a la Web» (Giraldo, 2005).

En PICSEL1, la base de conocimientos es descrita utilizando el lenguaje de representación del conocimiento CARIN (Levy et Rousset, 1998). Este lenguaje garantiza la «décidabilité»[4] (capacidad de decisión) del cálculo completo de lo que en PICSEL se llama «planes de consultas». Este trabajo sirvió de punto de partida para el trabajo que se efectuó en PICSEL2. Así el objetivo de PICSEL2 ha sido el proponer un método y unas técnicas automatizadas para ayudar a construir una ontología cuyo modelo corresponda al de aquélla de PICSEL1.

4 Un sistema lógico es «décidable» si y sólo si existe un algoritmo tal que por cada enunciado bien construido de ese sistema, exista un número finito N, tal que ese algoritmo pueda decidir en al menos N etapas algorítmicas si el enunciado es válido o es inválido.

El modelo de la ontología consiste en una jerarquía de clases que describen categorizaciones de clases de objetos del dominio de aplicación. En el caso de PICSEL1 el dominio fue el turismo. En este dominio, el modelo corresponde a una jerarquía que representa todo lo que puede venderse, en relación con el turismo, como: alojamientos, trayectos, localizaciones, etc. El modelo posee, además, otras jerarquías de clases que describen categorizaciones de conjuntos de objetos de subdominios del dominio de productos del turismo, como por ejemplo: lugares, pasatiempos, prestación de servicios, equipos de entretenimiento, etc. Los términos de la ontología de PICSEL1 están en francés. Cada clase es definida a través de las relaciones con otras clases. Para una clase dada, el modelo precisa la clase que la generaliza (clase madre en la jerarquía) y eventualmente sus propiedades específicas o bien el conjunto de propiedades necesarias y suficientes de un objeto para pertenecer a esa clase. Así, el esquema deseado en PICSEL2 es un esquema con base en clases que pueda ser representado en CARIN, el lenguaje de representación del conocimiento de PICSEL y, entonces, ser explotado por el conjunto de herramientas desarrolladas en el marco del proyecto, en particular, por el motor de consultas.

En PICSEL2 el interés está centrado en los sistemas mediadores que agrupan un conjunto importante de fuentes de información XML (eXtensible Markup Language) relativas a un mismo dominio de aplicación. ¿Por qué XML? Porque XML es desde hace algunos años, uno de los lenguajes más utilizados para describir y compartir información en la Web y por ello su consideración es esencial. XML es un lenguaje con base en etiquetas (*tags,* en inglés) que permite descubrir la estructura lógica de documentos, principalmente textuales (Michard, 1999). El etiquetaje estructural es un concepto simple que consiste en reconocer que todo documento textual está construido según una estructura. Esta estructura es explícita. La utilización de etiquetas permite marcar los elementos que componen esa estructura e indica, por lo tanto, el principio y el fin de cada uno de esos elementos. El documento de la figura 1 describe un recetario de cocina; allí algunos ejemplos de etiquetas son: Receta, Nombre, Descripción, Ingredientes, entre algunos.

FIGURA 1. EJEMPLO DE UN DOCUMENTO XML

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE receta SYSTEM "receta.dtd">
<Receta>
        <Nombre> "Carne de ternera encebollada" </Nombre>
        <Descripción>"Este exquisito plato se sirve con puré de papa o papas fritas" </Descripción>
        <Ingredientes>
            <Ingrediente>
                <Cantidad unidad= "Kg" >1</Cantidad>
                <Producto opcional= "no" esVegetariano= "no">"Cebollas"</Producto>
            </Ingrediente>
            <Ingrediente>
                <Cantidad unidad= "Kg" >1.5</Cantidad>
                <Producto opcional= "no" esVegetariano= "no">"Aleta de ternera" </Producto>
            </Ingrediente>
            <Ingrediente>
                <Cantidad unidad= "" >al gusto</Cantidad>
                <Producto opcional= "si" esVegetariano= "si">"Sal" </Producto>
            </Ingrediente>
            <Ingrediente>
                <Cantidad unidad= "vaso grande" >1</Cantidad>
                <Producto opcional= "no" esVegetariano= "si">"coñac"</Producto>
            </Ingrediente>
        </Ingredientes>
        <Instrucciones>
            <Paso>
              "Ponemos en una cazuela de barro un chorro de aceite y encima a rodajas el kilo de cebollas"
            </Paso>
            <Paso>
              "Ponemos encima la ternera y la salamos a gusto"
            </Paso>
            <Paso>
              "Ponemos la bandeja de barro en el horno a 250 grados y la asamos por los dos lados, a la vez que
              se asa añadimos el coñac."
            </Paso>
            <Paso>
              "Dejamos que se enfríe y trinchamos en filetitos la carne y la cebolla con el jugo que ha soltado y un
              poco de agua lo pasamos por la batidora, lo ponemos todo en una cacerola a que dé un hervor"
            </Paso>
        </Instrucciones>
</Receta>
```

En PICSEL2, las fuentes consideradas son, por lo tanto, homogéneas con respecto al formato de representación de los datos, puesto que todas son documentos XML. Al contrario, ellas son heterogéneas desde el punto de vista semántico[5], porque estos documentos pueden haber sido escritos por personas diferentes que no tienen los mismos criterios, ni para estructurar la información, ni para asignar los nombres de las etiquetas en los documentos.

En la presente investigación se propone una solución para la construcción automatizada de la ontología de los sistemas mediadores, ella explota la estructura de documentos XML que, en el marco

5 Relativo a la significación de las palabras.

de este proyecto, se suponen válidos y bien formados. Un documento es válido cuando obedece a una estructura de tipo predefinida, es decir, a una DTD (Document Type Definition), y es bien formado si respeta las reglas sintácticas de XML. El documento de la figura 1 es válido y bien formado (ver DTD figura 2).

Una DTD permite especificar qué elementos son susceptibles de aparecer o no en el documento, qué elementos pueden estar contenidos en otros elementos y en qué orden éstos deben aparecer. Lo más común es que la DTD esté definida en un archivo externo al documento XML. Este último, en su interior, hace referencia a la DTD a través de un URL (Unique Ressource Location).

FIGURA 2. EJEMPLO DE DTD EN FORMATO TEXTO

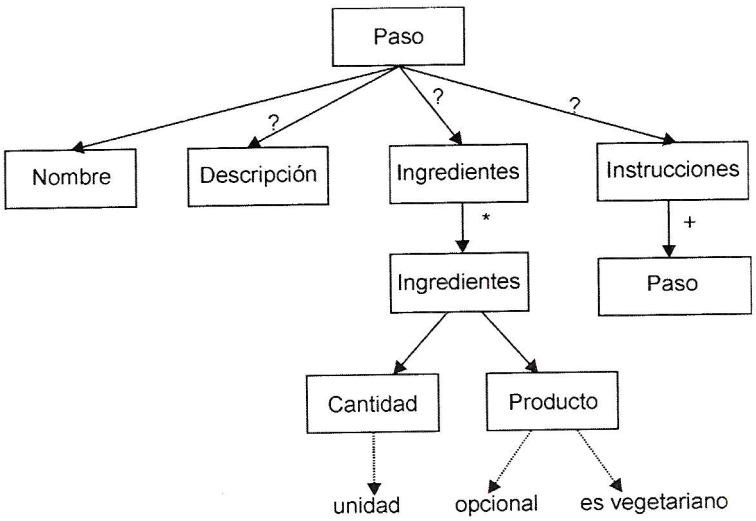
```
<!ELEMENT Receta (Nombre, Descripción?, Ingredientes?, Instrucciones?)>
<!ELEMENT Nombre (#PCDATA)>
<!ELEMENT Descripción (#PCDATA)>
<!ELEMENT Ingredientes (Ingrediente)*>
<!ELEMENT Ingrediente (Cantidad, Producto)>
<!ELEMENT Cantidad (#PCDATA)>
<!ATTLIST   Cantidad unidad CDATA #REQUIRED>
<!ELEMENT Producto (#PCDATA)>
<!ATTLIST   Producto opcional          CDATA "0"
                      esVegetariano  CDATA "si">
<!ELEMENT Instructiones (Paso)+>
<!ELEMENT Paso (#PCDATA)>
```

Una DTD es una representación abstracta de documentos XML (figura 2) que puede ser vista como un grafo orientado (figura 3), donde cada nodo del grafo corresponde a un elemento de la DTD, y una liga entre dos nodos corresponde a una liga de composición entre dos elementos (flecha continua), o a una liga entre un elemento y un atributo (flecha punteada). Los símbolos que se encuentran al lado de los términos Descripción, Ingredientes, Ingrediente, Paso, etc.[6], en la figura 2, son llamados caracteres de ocurrencia y tienen el siguiente significado:

6 Note que estos símbolos están también presentes sobre las ligas en la figura 2.

«?»: indica que el elemento puede aparecer 0 ó una vez
«*»: indica que puede aparecer 0 ó varias veces
«+»: indica que el elemento debe aparecer al menos una vez

FIGURA 3. EJEMPLO DE DTD EN FORMA DE GRAFO



3. SOLUCIÓN AL PROBLEMA DE LA CONSTRUCCIÓN DE LA ONTOLOGÍA

La solución propuesta explota prioritariamente las DTDs, en lugar de explotar los propios documentos XML, dado que se busca disminuir la masa de información por tratar para hacer más eficaz el proceso de construcción.

Una DTD es una representación asimilable a un esquema de un documento, en el sentido del esquema de una base de datos relacional. Es un modelo abstracto de uno o varios documentos XML. Los datos reales no se representan allí. Este nivel de representación se asemeja a una ontología, en tanto que esta última es una vista abstracta de un dominio. Sin embargo las DTDs no son descripciones de

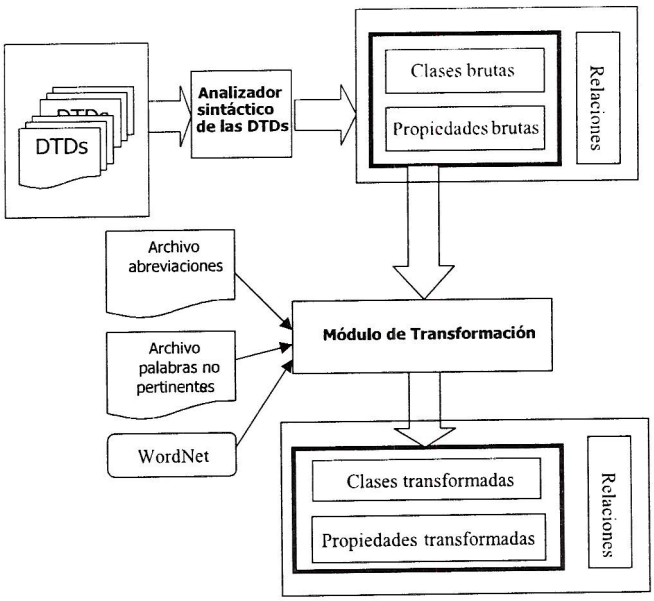
un dominio de aplicación y por ello la construcción de una ontología no puede ser un proceso de fusión de DTDs relativos a un dominio. Las DTDs no poseen una descripción explícita, ni de las clases de un dominio, ni de sus propiedades, ni de los diferentes tipos de relaciones entre las clases. La solución que se propone no consiste en fusionar DTDs sino en extraer las clases, las propiedades y las relaciones relativas a un cierto dominio, a partir de unas DTDs que no representan explícitamente esos elementos, pero que al menos han sido construidas por personas que tienen conocimiento del dominio y que lo han utilizado en el momento de representarlo. La solución que se propone comprende tres fases: la primera es una fase de extracción de componentes de la ontología: clases, propiedades y relaciones. La segunda fase organiza los elementos extraídos en la primera fase, y la tercera representa en CARIN, el lenguaje utilizado por el motor de consultas de PICSEL, el conocimiento anteriormente extraído y organizado. Por limitaciones de espacio, en el presente artículo sólo se presentan las dos primeras fases.

3.1 Fase de extracción de los componentes de la ontología

Esta fase es semi-automática. Aunque la extracción de los términos es completamente automática, el ingeniero constructor del sistema debe intervenir después de la extracción. El sistema debe permitirle actuar sobre los datos extraídos, con el fin de garantizar que los términos que van a ser parte de la ontología sean lo más pertinentes posible.

La muestra de datos a partir de la cual se realiza la etapa de extracción debe contener solamente DTDs relativas a un mismo dominio de aplicación y debe ser representativa de ese dominio. En ese sentido, la muestra debe corresponder a un número importante de DTDs que cubran el dominio estudiado. La noción de cobertura de un dominio no puede ser probada, por lo tanto, es primordial que las técnicas propuestas para construir la ontología sean lo suficientemente automatizadas, de tal manera que puedan ser, fácil y regularmente, reaplicadas a partir de una nueva muestra que cubra más ampliamente el dominio de estudio.

FIGURA 4. ARQUITECTURA DEL MÓDULO DE EXTRACCIÓN



El objetivo de esta fase es descubrir las clases del dominio, sus propiedades y las relaciones entre clases: relaciones de generalización/ especialización y relaciones específicas del dominio. Para extraer los componentes de la ontología, un módulo de extracción ha sido desarrollado. Su arquitectura es ilustrada en la figura 4. La extracción automática es realizada en dos tiempos. En un primer momento la extracción de términos es realizada. En un segundo momento, con el fin de obtener un conjunto de términos adaptados al uso que se desea hacer de la ontología, ciertos términos deben ser modificados. Las transformaciones efectuadas sobre los términos brutos son de diferente naturaleza; por ejemplo, es necesario tratar las abreviaciones, suprimir los términos dobles, lematizar[7] algunos términos, particularmente reemplazar los plurales por los singulares.

7 Lematizar un término consiste en darle la forma canónica correspondiente al término que existe en el diccionario.

En las subsecciones siguientes se explica el proceso de extracción propiamente dicho y luego el tratamiento que se debe hacer sobre los datos brutos extraídos de manera que se obtenga un vocabulario apropiado para la ontología; finalmente, la metodología propuesta y una experimentación realizada.

3.1.1 Análisis sintáctico de las DTDs

Es a través de un análisis sintáctico de DTDs, que las componentes de la ontología son extraídas. El principio de base para la fase de extracción es el siguiente: una clase es vista como una representación abstracta de un conjunto de objetos «complejos» los cuales son identificables en las DTDs, por el hecho de que se trata de elementos que se descomponen, es decir, son el objeto de una declaración de tipo ELEMENT que contiene uno o varios elementos hijos. Dado que nuestra solución reposa sobre la explotación de DTDs representativas del dominio, sus componentes deben aparecer *al menos* en una de las DTDs consideradas en la muestra de entrada a la fase de extracción. Siguiendo ese mismo razonamiento, se deduce un método de localización de los términos asociados a las propiedades: estos son los términos asociados a los elementos que *jamás* se descomponen en ninguna de las DTDs consideradas. El principio utilizado en la solución encuentra su justificación en el hecho de que las DTDs explotadas se suponen representativas del domino por cubrir. La implementación del principio que se ha descrito anteriormente, se efectúa aplicando las heurísticas siguientes:

- Heurística para determinar los términos de las DTDs que corresponden a clases:

  H_C: las clases son indicadas por los términos correspondientes a *al menos* un nodo no hoja en *al menos* un árbol de las DTDs consideradas como entrada.

- Heurística para determinar las propiedades:

  H_P: las propiedades son indicadas por los términos que corresponden siempre y únicamente a nodos hojas en *todos*

los árboles de las DTDs donde aparecen esos términos. De aquí en adelante llamaremos «términos-clases» y «términos-propiedades» a los términos que en las DTDs corresponden respectivamente a clases y a propiedades.

- Heurísticas para determinar los diferentes tipos de relaciones:

  H_{R1}: una declaración de tipo ELEMENT que describa la descomposición de un término-clase A en varios términos-clases B, C,... traduce una relación específica al dominio (R_{ED}), entre las clases A y B, A y C, etc.

  H_{R1}.: una expresión de tipo ELEMENT definiendo la composición de un término-clase A como una disyunción de otros términos-clases (B,C,...) traduce una relación específica al dominio (R_{ED}), siempre y cuando el indicador de ocurrencia exprese la posibilidad de una multivaluación (i.e * ó +).

  H_{R2}: una declaración de tipo ATTLIST definiendo los términos-propiedades P1, P2,... de un término-clase C traduce una relación de caracterización (R_C) entre la clase C y las propiedades P1, P2, etc.

  Otras relaciones de caracterización se pueden deducir de una declaración de tipo ELEMENT que traduce la descomposición de un término-clase C, conforme a la definición de la heurística H_C, en unos términos-propiedades, como se ha definido en la heurística H_P.

  H_{R3}: una expresión de tipo ELEMENT que define la composición de un término-clase A como una disyunción de otros términos-clases B, C,... traduce una relación de especialización (R_E), entre las clases B y A, C y A, etc. (B y C son clases más específicas que A), siempre y cuando el indicador de ocurrencia no exprese la posibilidad de una multivaluación (i.e. ? ó « »).

La extracción de clases y propiedades corresponde a una extracción de términos que indican respectivamente clases y propiedades. Sin embargo, la extracción de las relaciones no corresponde a la extrac-

ción de términos, a los cuales se les pueda asignar un nombre. El sistema de extracción identifica las relaciones por los términos que ellas ligan, pero no les da un nombre. Así la extracción de relaciones corresponde a una extracción de parejas de términos-clases o parejas formadas por un término-clase y un término-propiedad.

3.1.2 Tratamiento de los datos brutos extraídos

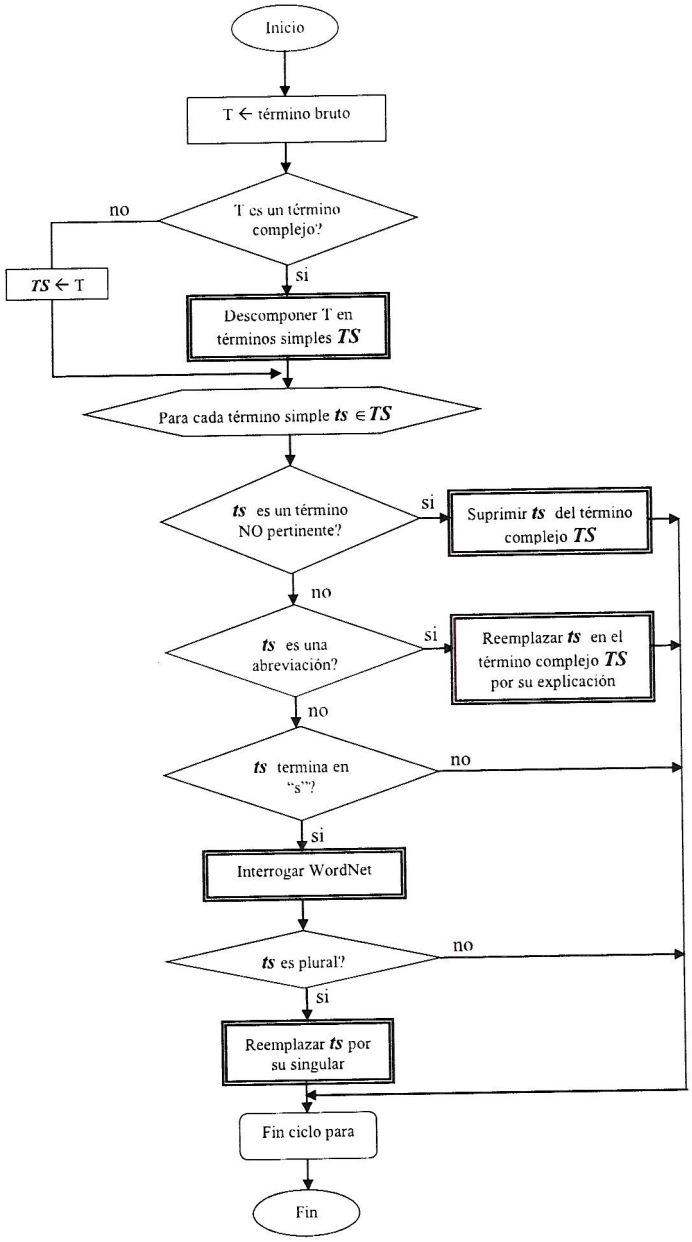
En esta subsección se presenta el módulo de transformación que permite tratar los datos brutos extraídos. El objetivo de este módulo es hacer unos tratamientos sobre los términos-clases y los términos-propiedades extraídos, eliminando los dobles, extrayendo las diferentes palabras de expresiones compuestas, remplazando las abreviaciones por su significado y eliminando ciertos términos no pertinentes de la ontología. No hay necesidad de tratar las relaciones, porque como se dijo anteriormente, a las relaciones no se les asigna un nombre. La figura 5 muestra el diagrama correspondiente al tratamiento de un término bruto.

Los problemas que se presentan son los siguientes:

- Los términos extraídos pueden corresponder a términos complejos. Se llama término complejo a aquel que está compuesto por varios términos simples. Un término simple es una palabra. Los términos simples incluidos en un término complejo son, en general, identificables por el hecho que ellos comienzan por una letra mayúscula o porque ellos son separados por un carácter especial (un guión, por ejemplo). FlightSegment es un término complejo compuesto por los términos simples Flight y Segment.
- Algunos términos extraídos corresponden a abreviaciones o acrónimos que es necesario explicar. Consideremos el término complejo AirSrvcClassPref[8], el cual contiene los términos

8 En el dominio del turismo este término indica la preferencia para el tipo de clase de servicio que se desea en un vuelo, por ejemplo: primera clase, clase ejecutiva, clase económica, etc.

FIGURA 5. TRATAMIENTO DE UN TÉRMINO BRUTO



simples Air, Srvc, Class y Pref. Se observa la presencia de dos abreviaciones: Srvc y Pref. Para explicar las abreviaciones se cuenta con un archivo que contiene para cada abreviación su explicación. El sistema busca en el archivo de abreviaciones todos los términos simples contenidos en un término complejo y si es encontrado es remplazado por su explicación. En nuestro ejemplo, el término complejo quedaría después del remplazo como AirServiceClass Preference.

- Los creadores de las DTDs, algunas veces, asignan a las etiquetas unos términos que no son significativos para una ontología que va a utilizarse en un contexto de mediación. Esos términos pueden tener algún significado para la persona que ha creado la DTD y para los que han creado los documentos instancias de esas DTDs. Sin embargo, esos términos pueden no hacer parte del vocabulario útil para la formulación de consultas. Estos términos los llamamos términos no pertinentes. De la misma manera que con las abreviaciones, el sistema utiliza un archivo de términos no pertinentes.

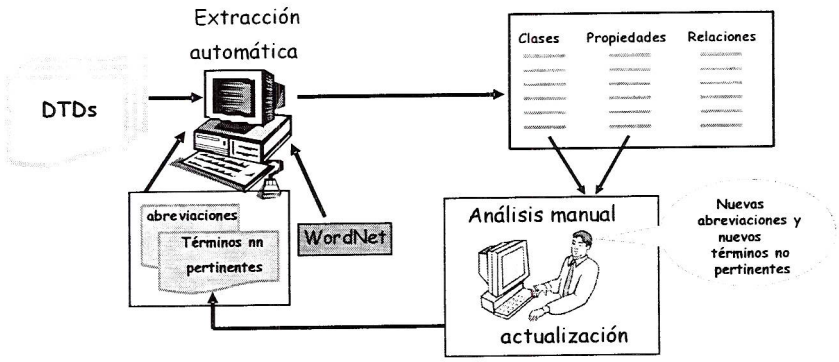
Finalmente, el sistema considera que algunos términos son diferentes simplemente porque el nombre (plural/singular) es diferente. Para detectar automáticamente que se trata de la misma palabra, el módulo de transformación explota el tesauro WordNet (Miller, 1995), que aunque no posee este servicio como una opción, hemos ideado una manera de utilizarlo para ello[9].

3.1.3 Metodología

La metodología propuesta para extraer los elementos que van a hacer parte de la ontología es interactiva e iterativa (figura 6). Es interactiva porque el ingeniero constructor del sistema juega

9 La manera como se utilizó WordNet para este efecto es explicada en Giraldo (2005), página 90.

FIGURA 6. EXTRACCIÓN DE COMPONENTES DE LA ONTOLOGÍA: PROCESO ITERATIVO E INTERACTIVO



un rol importante; e iterativa, porque es repetitiva. El ingeniero no tiene que ser necesariamente un experto, pero sí debe tener algún conocimiento del dominio de estudio.

Esta metodología posee dos etapas, la primera es de análisis macroscópico y la segunda de análisis *microscópico*. En la primera se hace una extracción automática de términos candidatos a ser parte de la ontología, aplicando las heurísticas a las DTDs. Así, esta etapa toma como entrada el conjunto de DTDs representativas del dominio de estudio y los archivos de abreviaciones y de términos no pertinentes del dominio. En un principio, estos dos últimos archivos, eventualmente, podrían estar vacíos. Después de la extracción automática, el ingeniero del sistema lee los términos extraídos, identifica tanto nuevos términos no pertinentes como nuevas abreviaciones y actualiza los archivos correspondientes.

Una vez que el análisis macroscópico ha sido hecho, el proceso de extracción puede ser reiterado hasta que el ingeniero del sistema esté satisfecho con el resultado obtenido. El objetivo de esta fase es doble: por una parte permite al ingeniero familiarizarse con el vocabulario del dominio y, por otra, filtra los términos que van a hacer parte de la ontología.

La segunda etapa de la metodología consiste en un análisis «micro», es decir, un análisis más profundo de los términos ex-

traídos. El ingeniero del sistema verifica si los términos extraídos están bien etiquetados como clases y como propiedades. Esto puede llevarlo a eliminar términos, a modificar ciertos nombres y a transformar ciertos términos-clases en términos-propiedades. Por ejemplo, aplicando las heurísticas, el sistema puede encontrar que el término *dirección* es una clase, ya que se descompone en *calle, número, barrio*, etc. Por el contrario, el ingeniero del sistema puede considerar que *dirección* es más bien una propiedad que puede caracterizar un inmueble. Además, él puede considerar que ciertos términos extraídos no son muy claros para el usuario o que no son términos del dominio, entonces puede cambiar el nombre o eliminarlos por considerarlos inútiles.

3.1.4 Experimentación

El dominio de los productos del turismo ha sido escogido como dominio de experimentación. Un prototipo del módulo de extracción ha sido desarrollado e implementado en el lenguaje Java. Él realiza la extracción de clases, propiedades y relaciones específicas al dominio (R_{ED}), de caracterización (R_C) y de especialización (R_E), a partir de un conjunto de DTDs y de esquemas XML. En las subsecciones siguientes se explica el origen de los datos, luego los resultados obtenidos y, finalmente, un análisis de los resultados.

3.1.4.1 Datos de entrada

Los datos de entrada a la experimentación corresponden a un conjunto de DTDs y esquemas XML[10] provistos por la OTA (*Open Travel Alliance* http://www.opentravel.org). La OTA es un consorcio que agrupa más de 150 empresas de la industria del turismo: agencias de viajes, hoteles, agencias de alquiler de carros, compañías aéreas, etc. En asociación con DISA (Data Interchange Standards Association), la OTA ha desarrollado unos estándares

10 Como las DTDs no satisficieron todas las necesidades inherentes a XML, surgieron los esquemas XML (XML Schemas) los cuales incorporaron mejoras, como por ejemplo el manejo de múltiples tipos de datos.

de comunicación basados en XML para facilitar el uso del comercio electrónico. Estos estándares han sido creados (en inglés) por los expertos de cada subsector del sector del turismo; por lo tanto, se espera encontrar en ellos el vocabulario propio a cada sector. Entre esos estándares encontramos 15 DTDs (317 líneas) relativas a las necesidades y preferencias de los viajeros, las cuales para efectos de la experimentación llamaremos serie 1, y 77 esquemas XML (5717 líneas) que complementan esta información, incluyendo reservación de vuelos, de hoteles, alquiler de automóviles. (Serie 2). En el momento que se inició esta investigación y se comenzó a desarrollar el analizador sintáctico, los esquemas XML todavía no habían sido aceptados como un estándar por el consorcio W3C[11], por lo tanto los esquemas XML de la OTA se han traducido en DTDs utilizando una herramienta de software (XML Spy).

Por otra parte, se construyeron dos archivos, uno de abreviaciones del turismo y otro con términos no pertinentes al domino. Estos archivos han sido iniciados con algunos términos, que se detectaron al dar una primera mirada a los datos de entrada.

3.1.4.2 Aplicación de las heurísticas a los datos de entrada

Gracias al analizador sintáctico (*parser*, en inglés) el proceso de extracción es realizado aplicando las heurísticas explicadas en la sección 3.1.1. Luego los términos brutos extraídos son tratados por el módulo de transformación (ver sección 3.1.2).

3.1.4.3 Resultados

Dado que los datos de la OTA están en inglés, los términos-clases y los términos-propiedades obtenidos como resultado de la fase de extracción están también en inglés.

---

11 «Word Wide Web Consortium» es una asociación internacional formada por organizaciones, personal y el público en general, que trabajan conjuntamente para desarrollar estándares Web. La misión del W3C consiste en guiar la Web hacia su máximo potencial mediante el desarrollo de protocolos y pautas que aseguren el crecimiento futuro de la Web.

a. Las clases y las propiedades

En la tabla 1 se presentan los resultados relativos a la extracción de clases y propiedades. Los términos-clases y los términos-propiedades han sido extraídos de manera completamente automática por el sistema, a partir de las dos series de datos (serie 1 y serie 2).

TABLA 1. NÚMERO DE TÉRMINOS EXTRAÍDOS POR CATEGORÍA

	Antes de tratamiento		Después de tratamiento	
	Clases	Propiedades	Clases	Propiedades
Serie 1	168	167	61	152
Serie 2	468	851	389	841

b. Las relaciones

La tabla 2 presenta el número de relaciones de especialización, de relaciones específicas al dominio y de relaciones de caracterización extraídas de manera automática a partir de las dos series de datos de la OTA.

TABLA 2. NÚMERO DE RELACIONES EXTRAÍDAS

	Relaciones especialización	Relaciones específicas al dominio	Relaciones de caracterización	Número total de relaciones extraídas
Serie 1	2	55	343	400
Serie 2	8	386	1.309	1.703

A título de ilustración, la tabla 3 muestra algunos ejemplos de términos ligados por las diferentes categorías de relaciones.

Las relaciones de especialización (R_E) indicadas en la tabla 3 fueron extraídas a partir de la serie 2. En realidad muy pocas relaciones de este tipo fueron encontradas (sólo 8). En la tabla sólo se presentan 6, pues dos de ellas no pertenecen al dominio de estudio (el turismo).

TABLA 3. EJEMPLO DE RELACIONES EXTRAÍDAS AUTOMÁTICAMENTE

Tipo de relación	Términos ligados
Relación de especialización	R_E (*PlanType*) = { *Annual, Package* } **Explicación**: existen dos tipos de seguros : anuales y paquete (el seguro de tipo « package » tiene que ver con todos los elementos de un viaje, tales como transporte, alojamiento, alquiler de vehículos, servicios especiales,...). R_E (*GuaranteeInformation*) = { *GuaranteeDeposit* } **Explicación**: Una caución es una información ligada a las garantías. R_E (HotelSearchValue) = { Area, HotelReference, ReferencePoint } **Explicación:** La zona geográfica, la referencia a un tipo de hotel (hotel de una cadena particular por ejemplo) y la proximidad a un punto de referencia son criterios de búsqueda de un hotel.
Relación especifica al dominio	R_{ED} (*AirBook*) = { (*AirItinerary*, "") } **Explicación:** La reservación de un vuelo tiene relación con un cierto trayecto aéreo R_{ED} (*AirItinerary*) = { (*FlightSegment*, +) } **Explicación:** Un trayecto aéreo puede estar compuesto de diferentes vuelos
Relación de caracterización	R_C (*FlightSegment*) = { (*DepartureAirport*, "")*,*(*JourneyDuration*, ?) } **Explicación:** El aeropuerto de salida y la duración del viaje son características de un vuelo.

3.1.4.4 Análisis

Este acercamiento a la solución ha permitido diferenciar bien los términos correspondientes a clases (aprox. 30%) y aquellos correspondientes a términos propiedades (tabla 4). Esto es importante, dado que el modelo de la ontología que se desea construir es un modelo con base en clases; es decir, su estructuración se efectúa en torno a las clases del dominio. Así, este acercamiento permite reducir considerablemente el número de términos por considerar en el momento de la construcción de la jerarquía de clases.

Los tratamientos efectuados en los términos brutos extraídos, bien, si ellos son simples, han sido bastante efectivos. En efecto, el conjunto de términos que son eliminados por medio del tratamiento son pocos: en la serie 1, antes 68 y después 61; en la serie 2, antes 468 y después 389 (tabla 5).

TABLA 4. TÉRMINOS-CLASES Y TÉRMINOS-PROPIEDADES EXTRAÍDOS EN NÚMERO Y PORCENTAJE

Nº serie	Términos-clases (número, %)	Términos-propiedades (número, %)
1	61 términos 31/213 = 28%	152 términos 152/213 = 71%
2	389 términos 389/1230 = 31%	841 términos 841/1230 = 68%

TABLA 5. IMPACTO DEL TRATAMIENTO SOBRE LOS DATOS BRUTOS EXTRAÍDOS

Serie	Número de términos-clases antes del tratamiento	Número de términos-clases después del tratamiento	Términos-clases modificados por el tratamiento	Términos-clases no modificados por el tratamiento
1	68	61	47 (47/68=69%)	21 (21/68=31%)
2	468	389	396 (396/468=84%)	72 (72/468=16%)

Ciertos tratamientos efectuados en los términos brutos extraídos reposan en la utilización de un tesauro en línea, el cual es bastante eficaz. La interrogación de WordNet es muy útil para reconocer los términos en plural e indicar el término singular correspondiente. Así en la serie 2, el 22% de los términos-clases brutos han sido identificados por WordNet como plurales. Dado que al convertir un término que se encuentra en plural en su equivalente singular, el resultante puede ser un término ya existente entre los términos extraídos, por lo tanto es considerado sólo una vez.

Como parte de esta experimentación se realizó un pequeño ejercicio que consistió en hacer una extracción manual de términos-clases y de términos-propiedades, a partir de los datos de la serie 1 y compararlos con los extraídos automáticamente. El resultado fue el siguiente: el 79% de los términos fueron extraídos por los dos métodos. Dado que partimos de la hipótesis que los términos extraídos manualmente son los correctos, podríamos decir que el proceso automático ha «olvidado» extraer aproximadamente el 21% de los términos. Esto parece preocupante, pero en realidad

lo que sucedió fue lo siguiente: los términos-clases «olvidados», sí aparecían extraídos por el método manual pero con un nombre diferente. En efecto, el ingeniero del sistema en el momento de la extracción manual interpretaba el significado de las clases extraídas de las DTDs y deducía los términos-clases que le parecían más «apropiados» para hacer parte de la ontología del dominio. Pero el proceso automático no tiene toda esa capacidad de interpretación. Así, por ejemplo, el término RelatedTraveler hace parte de los términos extraídos automáticamente, mientras que el ingeniero prefiere cambiar el nombre por Traveler. El mismo fenómeno se observó con respecto a los términos-propiedades. Por lo tanto, el método de construcción propuesto es interactivo e iterativo, como se explicó en la sección 3.1.3.

Una experiencia de construcción automática de una ontología fue llevada a cabo en el proyecto ASIUM (Acquisition of SemantIc knowledge Using Machine learning methods), (Faure et Nédellec, 1999). Este proyecto utiliza técnicas de la inteligencia artificial, como es el aprendizaje automático, para extraer los términos de la ontología a partir del análisis sintáctico de textos técnicos. Al igual que nosotros, los investigadores que realizaron este proyecto advierten la dificultad de una generación «completamente automática» de ontologías y la necesidad de la intervención humana para resolver las ambigüedades en algunos conceptos y relaciones.

3.2 Fase de estructuración de los elementos extraídos

Esta es la fase que sigue a la fase de extracción. La fase de estructuración toma como entrada la salida de la fase de extracción, es decir, el conjunto de clases, de propiedades y de relaciones. El objetivo de esta fase es organizar los elementos extraídos en una estructura compuesta de un conjunto de términos relacionados conforme con el modelo definido en la sección 2.

Proponer una organización tal, no es inmediato. Puesto que es muy probable que entre un conjunto de DTDs múltiples y heterogéneas se encuentren también múltiples formas de estructurar estos documentos y, por lo tanto, las relaciones de especialización

van a reflejar esas múltiples formas. Frente a esas múltiples escogencias (a veces incompatibles) es difícil determinar de manera automática la estructura de la ontología que se desea construir. La solución propuesta parte de un bosquejo de ontología de clases simple, aplicado al dominio de estudio. La jerarquía de clases se dice simple en la medida en que ella no posee muchos niveles. Se propone al ingeniero creador del sistema definir los dos primeros niveles de la jerarquía. Esta fracción de jerarquía sirve de punto de partida para la organización del conjunto de clases extraídas en la primera fase. Una vez construidos esos dos primeros niveles se busca, entre el conjunto de clases extraídas, ésas que deben aparecer en la jerarquía inicial con el objetivo de completarla.

En el desarrollo de esta investigación se vislumbran dos soluciones para estructurar las clases extraídas a partir de las DTDs. La primera es la utilización de herramientas lingüísticas, y la segunda consiste en la utilización de datos estandarizados.

3.2.1 Utilización de herramientas lingüísticas

La jerarquía inicial es enriquecida explotando las relaciones de especialización extraídas de la muestra de DTDs. El problema de estructuración corresponde entonces a un problema de fusión de jerarquías parciales, extraídas, semiautomáticamente, con el bosquejo de ontología de clases, construido manualmente. Los tesauros lingüísticos provistos de relaciones semánticas entre términos pueden servir para facilitar esta fusión.

Para encontrar las similitudes sintácticas entre los términos-clases extraídos y los términos-clases de la jerarquía inicial es posible explotar las relaciones semánticas siguientes:

Igualdad: Dos términos X y Y son iguales si ellos se escriben de la misma manera.

Sinonimia: Dos términos X y Y son sinónimos si ellos tienen el mismo sentido

Hiponimia e hiperonimia: Si el término X es un «tipo de» término Y, X es una hiponimia de Y y Y es una hiperonimia de X.

WordNet ha sido utilizado para encontrar esas similitudes semánticas (excepto, evidentemente para la igualdad). WordNet es una base de datos lexical para el inglés que reagrupa sustantivos, verbos, adjetivos y adverbios, organizados en conjuntos de sinónimos («synsets»). Los conjuntos son ligados entre ellos por relaciones semánticas. En WordNet, una palabra puede tener varios sentidos. Cada sentido se encuentra en un «synset» diferente.

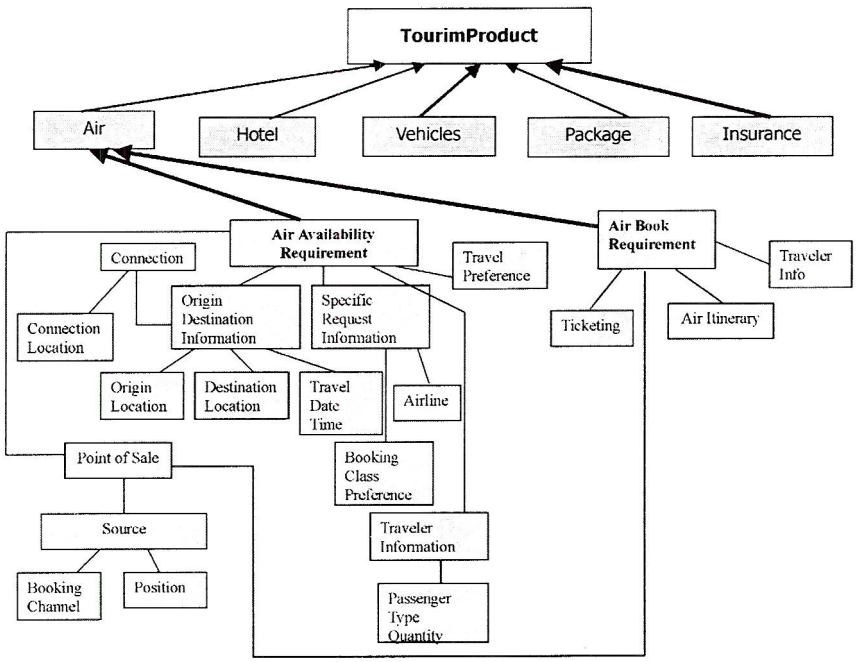
La idea intuitiva del proceso de enlace consiste en encontrar el buen lugar en la jerarquía inicial para colocar las clases de la jerarquía parcial, esto con la ayuda de las relaciones semánticas provistas por WordNet, considerando además las nuevas clases propuestas por WordNet. Así, al final del proceso de enlace, la jerarquía inicial será enriquecida con las clases de la jerarquía parcial y eventualmente, con otras clases.

3.2.2 Utilización de datos estandarizados

Ahora más que nunca con la llegada de la Web, los diferentes sectores han sentido la necesidad de desarrollar estándares que les permitan comunicarse entre ellos. Por ello, encontramos cada vez más, organismos cuyo objetivo es crear estándares. La solución que se propone explota esta situación construyendo una ontología a partir de DTDs desarrollados por un organismo de estandarización, la OTA. Como se ha dicho en la sección 3.1.4.1, la OTA es un organismo que agrupa varios sectores de la industria del turismo. Cada sector ha desarrollado un conjunto de DTDs, obedeciendo a ciertas convenciones predefinidas por dicho organismo.

Esta solución simplifica considerablemente la fase de estructuración. El proceso de enriquecimiento de la jerarquía inicial no requiere la utilización de tesauros. En efecto, la clase de nivel 1 de la jerarquía corresponde al nombre del dominio, en nuestro caso, el turismo. La etiqueta raíz de cada DTD da origen a una clase de nivel 2 en la jerarquía inicial. Todos los términos-clases extraídos de manera semiautomática forman una red de relaciones que son ligadas a una clase de nivel 2. La figura 7 ilustra un extracto de

FIGURA 7. EXTRACTO DE LA ONTOLOGÍA CONSTRUIDA A PARTIR DE DTDS DE LA OTA



la jerarquía de clases construida semiautomáticamente a partir de DTDs de la OTA.

Dado que los creadores de las DTDs son expertos del dominio, se espera que los términos utilizados para denominar las etiquetas sean los más apropiados y estén normalizados; por lo tanto, no habrá ni polisemia ni sinonimia.

Recordemos que si un término-clase es definido de diferentes maneras en varias DTDs, él aparecerá sólo una vez en la ontología y será descrito de acuerdo con las heurísticas definidas en 3.1.1. Esta última solución de estructuración fue implementada en el sistema OntoMedia.

4. Conclusión

Se ha presentado en este artículo una solución para la construcción semiautomática de una ontología (del dominio) a partir de un conjunto de DTDs representativas del dominio de aplicación. Una ontología construida a partir de tales datos, provee un vocabulario pivote, clave de la solución propuesta. La ontología es construida de manera cooperativa, solicitando la ayuda del ingeniero del sistema. Se ha mostrado como construir la ontología a partir de cualquier conjunto de DTDs, aplicando una serie de heurísticas para extraer las clases, las propiedades y las relaciones. Para estructurar estos elementos, dos soluciones han sido propuestas. La primera utiliza un tesauro para su estructuración. La segunda se da en un contexto particular: los estándares. Esta última consiste en explotar las DTDs construidas por un grupo de expertos del dominio en aras de crear unos estándares para mejorar la comunicación entre los diferentes sectores de un dominio dado.

La solución propuesta es general, dado que ella puede ser aplicada a partir de un conjunto de DTDs desarrolladas por cualquier organismo de estandarización. La construcción semiautomatizada de la ontología para un sistema mediador es un gran aporte hacia la automatización de sistemas mediadores, dado que la ontología corresponde a la parte del mediador que es específica al dominio. El motor de consultas de un sistema mediador, que es genérico, sólo se construye una vez, pero como la ontología es específica al dominio, hay necesidad de construir una para cada dominio de estudio. La construcción manual de cada una de ellas es tediosa y poco práctica.

5. Perspectivas

Una perspectiva importante de este trabajo se sitúa en el contexto de la utilización de las ontologías como un recurso útil para dar significado a la información que se encuentra en la Web. La expresión «Web Semántica» es utilizada inicialmente por Tim Berners-Lee, director actual de la W3C. La nueva generación Web

–La Web Semántica– tiene por objetivo resolver ese problema, es decir, hacer que la semántica de la información de la Web sea a la vez comprensible por los usuarios y por las entidades informáticas (motores de búsqueda, servidores de información, etc.). Para ello es necesario representar semánticamente el contenido de la Web. Hoy en día las ontologías son consideradas como una tecnología imprescindible para alcanzar la Web Semántica. Por ello, actualmente se está formulando una sublínea de investigación en el Instituto Tecnológico Metropolitano, que se ha denominado «Integración de información en el contexto de la Web Semántica».

Las ontologías juegan un rol central en el etiquetaje semántico de la información de la Web. Así, un tipo de ontología característico de la Web debe poseer una taxonomía y un conjunto de reglas de inferencia. Esta taxonomía, con el fin de representar las clases de objetos de un dominio y las relaciones entre ellas. Las clases, las subclases y las relaciones entre las entidades son una herramienta muy potente para utilizar mejor la Web. Se pueden expresar un gran número de relaciones entre las entidades, atribuyendo propiedades a las clases y permitiendo que las subclases hereden esas propiedades. Las reglas de inferencia en las ontologías son todavía más potentes. Una ontología puede expresar la regla siguiente: «si un código postal de una ciudad está asociado a un código de departamento y si una dirección utiliza ese código postal, entonces esa dirección está asociada al código de departamento». En consecuencia, un programa podrá deducir que la dirección del «INSTITUTO TECNOLÓGICO METROPOLITANO» situado en «Medellín», debe encontrarse en el «departamento de Antioquia», en «Colombia» y deberá por lo tanto formatear la dirección siguiendo los estándares adoptados por este país.

Es claro que muchos problemas persisten, por ejemplo, la confusión que se presenta si una ontología define el concepto «código postal» como parte de una dirección y otra ontología, queriendo significar lo mismo, lo define como «código ZIP». Es en estos casos donde un mecanismo de definición de relaciones de equivalencia entre conceptos de diferentes ontologías, es necesario. En nuestro

ejemplo: código postal equivalente a código ZIP. Este mecanismo puede ser implementado, utilizando igualmente ontologías.

Las ontologías pueden mejorar el funcionamiento de la Web de varias maneras:

- Se pueden utilizar para mejorar la pertinencia de las búsquedas, haciendo que el programa de búsqueda encuentre solamente las páginas que hacen referencia a un concepto preciso, en lugar de aquéllas que utilizan palabras claves ambiguas.
- Otras aplicaciones más avanzadas pueden utilizar las ontologías para asociar la información de una página Web a ciertas estructuras de conocimiento y a unas reglas de inferencia. Por ejemplo, la página Web del doctor «Elkin Patarroyo» –supóngase que ella posee una liga hipertexto hacia su bibliografía–, donde quien observe esta información podrá leer que él ha recibido su título de doctor en la Universidad de Rockefeller. Un programa de computador tendría que ser muy sofisticado para «adivinar» que la información allí encontrada es una bibliografía. Si esta página fuera debidamente etiquetada con los conceptos de una ontología, un programa de computador podría encontrar, por ejemplo, que el doctor Patarroyo recibió su título en «Rockefeller University», que es miembro de un proyecto de investigación particular, etc. Todas estas informaciones podrían ser tratadas instantáneamente por un computador y ser utilizadas para responder preguntas, como: «¿Dónde recibió el diploma de doctorado el señor Elkin Patarroyo?». Con los motores de búsqueda actuales, la respuesta a esta pregunta debe ser encontrada por el usuario, leyendo cuidadosamente su página.

Además, el etiquetaje de las páginas Web con los conceptos de ontologías, permite resolver preguntas complicadas cuya respuesta no se encuentra en una sola página, sino que hay que visitar varias páginas para encontrar la respuesta apropiada.

Es evidente, entonces, la importancia de esta sublínea de investigación que está naciendo en el ITM. Como toda línea de investigación que nace, requiere de la generación de una masa crítica que permita

crear proyectos en torno a esta temática. Actualmente en Colombia pocos investigadores trabajan en este campo[12]. El ITM, en busca de su excelencia académica, se propone investigar en temas que estén a la vanguardia del desarrollo mundial. Este artículo pretende ser, entonces, un comienzo para avanzar en esta dirección.

6. Bibliografía

BORST, W. N. (1997). «Construction of Engineering Ontologies», PhD Thesis, University of Twente, EnschedeCHAUDHURI S., DAYAL U. (1997) «An overview of Data Warehousing and OLAP Technology», En: SIGMOND Record, vol. 26, N°1, p. 65-74.

GIRALDO, G. L. (2005). «Construction automatisée de l'ontologie de systèmes médiateurs: application à des systèmes intégrant des services standards accessibles via le Web», Tesis doctoral, Université Paris Sud XI, Orsay, Francia.

GOASDOUE, F. (2001). «Réécriture de requêtes en termes de vues dans CARIN et intégration de informations». Tesis doctoral, Université Paris Sud XI, Orsay, Francia.

GRUBER. T. R. (1993). A translation Approach to Portable Ontology Specifications, En: Knowledge Acquisition, 5(2), p. 199-220.

GUARINON., GIARETTA P. (1995). «Ontologies and Knowledge Bases: Towards a Terminological Clarification». En: Mars N. J. I. (edit.). Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, p. 25-32.

LEVY A., ROUSSET M-C. (1998). «Combining Horn rules and description logics in CARIN». En: Artificial Intelligence Journal. Vol. 104, septiembre, p.165-209.

MICHARD A. (1999). XML langage et applications, Paris: Ed. Eyrolles.

MILLER G. A. (1995). «WordNet: A Lexical Database for English». En: Communications of the ACM, vol. 38, N°11, noviembre, p. 39-45.

12 Por citar algún grupo, en la Universidad del Valle (Cali-Colombia) existe I+DeaSWeb (Investigación y Desarrollo para la Semántica de la Web).

NECHES R., FIKES R., FININ T., GRUBER T., PATIL R., SENATOR T., SWARTOUT W. R. (1991). «Enabling Technology for knowledge Sharing». En: AI Magazine, vol. 12, p. 36-56.

REYNAUD Ch., GIRALDO G. (2003). «An application of the mediator approach to services over the Web». En: Data Integration in Engineering, Concurrent Engineering (CE'2003) - the vision for the Future Generation in Research and Applications. R. Jardim-Gonçalvez et al. (edit.), Portugal, vol.1, julio, p. 209-215, ISBN 90 5809 622 X.

WIEDERHOLD G. (1992). «Mediators in the architecture of future information systems». En: IEEE Computers. vol. 25, N°3, marzo, p. 38-49.