Generación semiautomática de ontologías utilizando bases de datos relacionales como fuente primaria de información

David González-Marrón[1,2], Miguel González-Mendoza[2] y Neil Hernández-Gress[2]

[1] Instituto Tecnológico de Pachuca,Carretera México-Pachuca Km 81.5, Pachuca, Hidalgo, México
dgonzalez@itpachuca.edu.mx

[2] Tecnológico de Monterrey, Carretera Lago de Guadalupe Km 2.5, Atizapán de Zaragoza, Edo. de México, México
{mgonza,ngress}@itesm.mx

Resumen Actualmente, existen muchos investigadores buscando automatizar la producción de ontologías, utilizando ingeniería inversa para la extracción de información contenida en las bases de datos, sin embargo debido a la continua evolución de estándares en las bases de datos relacionales, este proceso solo se ha podido realizar de manera parcial. Es por esto que muchas de las características en la definición de almacenamiento existentes en los modernos manejadores de bases de datos no son soportadas aún por la tecnología semántica. Otro factor importante es que una misma base de datos puede ser implementada con diferentes comandos de definición de datos por diferentes diseñadores, lo que ocasiona diferencias en las ontologías generadas. Este trabajo está enfocado principalmente a describir como pueden ser generadas las ontologías a partir de las bases de datos relacionales, los principales problemas para realizar esta actividad de manera automática, los lenguajes que pueden ser utilizados para realizar ésta actividad, y propone una alternativa para generar ontologías de manera semiautomática utilizando software libre desarrollado por la comunidad.

Palabras clave: Bases de datos heterogéneas, Web semántica, ontologías.

1. Introducción

La mayor parte de la información accesible en la web, es extraída de bases de datos utilizando programas especialmente diseñados. La cantidad de la información existente en estas bases de datos es más del 70 % de la información en la Web, esta información existente en las bases de datos es comúnmente conocida como la "deep web" [1] debido a que no es accesible a través de motores de búsqueda de propósito general. La web semántica está en búsqueda de mecanismos que permitan encontrar, compartir y combinar la información de la web más fácilmente [2].

EL uso de ontologías permiten incrementar la interoperabilidad entre los sistemas, al mismo tiempo permite el uso de la tecnología semántica para incrementar el po-

tencial de las consultas formuladas. En este trabajo un análisis de los mecanismos requeridos para producir ontologías automáticas es realizado, se mencionan los lenguajes y comandos mas utilizados para producir ontologías, adicionalmente un análisis de las ventajas y desventajas de explotación de información semántica para un usuario común es realizado, así como del proceso para producir ontologías. La web semántica busca la integración de archivos RDF, sin embargo debido a que la mayor parte de la información está almacenada en Bases de Datos, un mecanismo para interactuar con esta información debe ser implementado, uno de los métodos más utilizados por investigadores y desarrolladores es el mapeo de esquemas definidos en bases de datos en SQL a archivos en lenguajes OWL ó RDFS principalmente. En la siguiente tabla se muestra una relación entre SQL y la web semántica. Puede ser visto en la tabla en su nivel inferior que los datos pueden ser convertidos en archivos RDF (conformado por tripletas), en el nivel superior los triggers permiten describir las operaciones que se realizan al hacer operaciones de actualización de datos. Actualmente no existe una herramienta automática que genere reglas, principalmente por la complejidad de los triggers (disparadores). El estado del arte es la generación de ontologías básicas. Estas ontologías son nombradas por Sequeda et al [3] como "ontologías putativas". Debido a que son ontologías simples que requieren ser validadas y complementadas por expertos de dominio.

Tabla 1. Niveles de equivalencia entre componentes de SQL y lenguajes web semánticos.

SQL	Semantic Web
TRIGGERS	RULES
CONSTRAINTS	OWL
TABLE DEFINITION	RDFS
RELATIONAL MODEL	RDF

1.1. Integración de bases de datos

La integración de bases de datos, es considerada como una de las tareas más difíciles en la web semántica principalmente por el gran número de variables involucradas en esta actividad, tales como: incompatibilidad de datos, distribución de datos, diferente significado semántico de los datos, mecanismos de seguridad, tareas de mapeo de las bases de datos, diferente representación y precisión de los datos. Este trabajo está más enfocado a lograr interoperabilidad en las bases de datos utilizando ontologías utilizando un enfoque semántico. Los avances más significativos en el área han sido realizados por el grupo OMG [11]. Un importante estudio de las principales herramientas existentes para generar mapeos automáticos es realizado por el grupo incubador del RDB2RDF [12].

1.2. Aproximación para accesar BD relacionales utilizando métodos semánticos

Para crear mapeos utilizando bases de datos relacionales, la primera opción es la creación de mapeos estáticos del esquema de las BD, este proceso se encuentra aún en desarrollo por diferentes investigadores, el segundo es un proceso semiautomático que

transforma la información a un formato semántico. A continuación se hace una breve descripción de cada método

Realización de mapeos estáticos. En este método, el mapeo es realizado sobre todo el esquema de la BD sin realizar copias físicas de la información, la estructura de las tablas de las BD relacionales y sus atributos son mapeados a sus correspondientes clases y propiedades en una ontología definida por el usuario. El esquema de BD relacionales es una tupla R (U, D, dom, I). Donde : R es conocido como una relación definida como un conjunto de tuplas que tienen las mismos atributos , y una tupla usualmente representa información acerca de un objeto; U es un conjunto finito de atributos en una relación; D es el dominio de la BD; dom representa una función que mapea U a D , donde cada atributo Ai en U debe tener un valor válido en Di; I es un conjunto finito de restricciones de integridad, el cual restringe las instancias de datos almacenadas en la BD, dos diferentes tipos de integridad pueden ser aplicados (integridad de datos e integridad referencial). La siguiente figura muestra el proceso requerido para generar información a partir de una BD relacional que sea de utilidad para un motor de consultas semántico. Actualmente los resultados producidos por la tecnología no son satisfactorios, debido a que todavía se requiere un proceso de refinación por expertos en las áreas del dominio de discurso

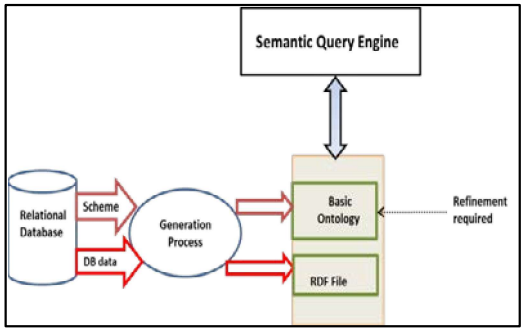

Fig. 1. Mapeo directo de bases de datos relacionales a datos semánticos.

Realización de conversión de las bases de datos (Wrapping Systems). Este método produce una conversión de la BD a un formato diferente, comúnmente conformado por tripletas de los siguientes tres elementos (Sujeto, Predicado y Objeto). Aquí la conversión produce un gran incremento de la información. Esta conversión generalmente es realizada en formato RDF, el cual es almacenado en repositorios que pueden ser explotados utilizando SPARQL. Diferentes herramientas existen para almacenar los archivos RDF generados de manera que los datos sean almacenados en una sola plataforma. EL proceso de creación de ontologías normalmente utiliza una ontología de dominio, la cual ayuda a producir información semántica, esta ontología de domi-nio si no es utilizada, produce normalmente tripletas en un formato común de RDF, en este caso se considera la premisa de un mundo cerrado “closed world assumption”, y los URIs “Uniform Resource Identifier” (identificadores de recursos uniformes) creados son muy simples, en cambio cuando se utiliza una ontología de dominio las URIs generadas incluyen información acerca del dominio utilizado.

2. Porqué utilizar la integración semántica

Los aspectos generales relacionados con las aplicaciones de las BD es que la integración semántica puede solucionar diferentes problemas que se presentan en las representaciones estructuradas, como la codificación de datos con más de una representación, es por eso que las aplicaciones deben resolver heterogeneidades con respecto a los esquemas y sus datos, la representación semántica permite la manipulación de datos y permite la transformación de datos entre diferentes esquemas de BD [13]. Esta tarea se relaciona principalmente con la conversión de datos para la explotación de consultas. Una conceptualización de explotación de consultas se muestra en la figura 2, en la cual un usuario solicita una consulta que requiere la integración de diferentes bases de datos, como la mayor parte de las veces los datos no son compatibles, el uso de la integración semántica es una propuesta a ser considerada.

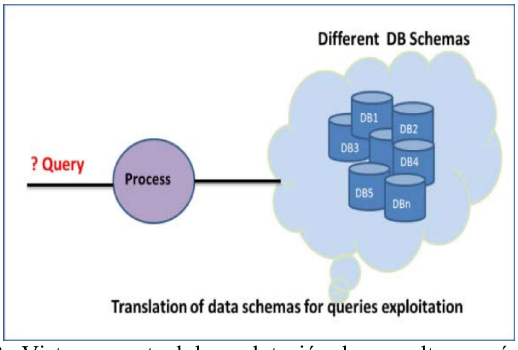

Fig. 2. Vista conceptual de explotación de consultas semánticas.

2.1. Alternativas para la integración de datos

Dos alternativas existen para lograr la integración de datos, la más común es la estática, utilizando ETL "Extracción, Transformación y Cargado (*Load*)", la cual produce conversión de los datos de un área temporal a un área de representación final. La otra aproximación es la dinámica (virtualización de datos), en la cual los datos permanecen en su fuente de origen, y una vista conceptual de datos es producida tan pronto se necesita. El ETL tradicional es una variación del paradigma ETL, en el cual las fuentes de datos son usadas para soportar aplicaciones de negocios, y estas aplicaciones normalmente requieren formatos diferentes, archivos, estructuras y aún diferentes codificaciones. Normalmente existe una tendencia para normalizar datos antes de integrarlos a un sistema objetivo.

La virtualización de datos es contraria al método tradicional de extracción de datos de diferentes fuentes y de almacenamiento temporal de estos datos en un área de trabajo. Este método denominado federación de datos o virtualización de datos permite que las fuentes de datos permanezcan en sus localizaciones originales. La virtualización es soportada por una variedad de fuentes de datos nativas y proporciona vistas relacionales sin requerir que los datos sean extraídos de su fuente. Los consumidores de datos toman una vista de datos estructural y semánticamente consistente.

Otra aproximación novedosa a ser utilizada para integrar datos es la integración semántica, la cual es un área activa dentro del área de las bases de datos, integración de información y ontologías, permitiendo la interoperabilidad entre diferentes sistemas. La integración semántica involucra también técnicas de alineación con esquemas de bases de datos para responder consultas complejas que requieren la utilización de múltiples fuentes, aquí los datos son transformados a fin de poder utilizar la semántica de los datos, dos diferentes alternativas son de uso potencial. El primero es similar a la aproximación utilizada por ETL, pero utilizando una organización semántica de los datos y replicándolos. El segundo está relacionado con la virtualización de los datos, creando archivos de mapeo necesarios para localizar los datos.

2.2. Porqué la integración semántica

La necesidad de uso de información semántica, es causada debido a la diversidad de representación de datos y la información que normalmente se encuentra interrelacionada entre si, es importante mencionar que debido a que la información no se encuentra contextualizada, la mayor parte de los desarrolladores podrían no pensar en los datos en si, sino en la estructura de los datos, tal como: tipos de datos, esquemas, formatos de archivos, construcciones de BD y otras estructuras que no están relacionadas directamente con la información.

3. Problemática para modelar ontologías usando bases de datos relacionales

Al momento de hacer conversiones de datos relacionales a datos semánticos, diferentes problemas se presentan, entre ellos la posibilidad de definir de diferentes maneras las llaves primarias y foráneas, para la realización de conversión de datos relacionales a semánticos, diez diferentes combinaciones básicas de estas llaves deben ser consideradas al momento de realizar un mapeo directo, en la tabla 2 se muestra ésta combinación básica. De igual manera la evolución que ha sufrido el lenguaje SQL como puede ser visto en la tabla 2, ha incrementado la complejidad para procesar e interpretar de una manera simple la información almacenada en las bases de datos y convertirla a información semántica

3.1. Diferentes comandos utilizados por los diseñadores, producen diseños de bases de datos equivalentes, pero sintácticamente diferentes en el lenguaje SQL-DDL

Los elementos más importantes para analizar el esquema de las bases de datos son las llaves primarias y foráneas, estas llaves a su vez pueden ser simples o compuestas. Las diferentes combinaciones que se pueden presentar son mostradas en la siguiente tabla, Un análisis mas profundo debe ser realizado en cada caso a fin de convertir correctamente el esquema definido en información semántica.

Tabla 2. Consideraciones para producir datos semánticos usando el esquema relacional.

Combinación básica de llaves primarias y foráneas existentes en una base de datos relacional
PK: Una relación que contiene únicamente una llave primaria
C-PK: Una relación que contiene una sola llave primaria compuesta
S-FK: Una relación que contiene únicamente una llave foránea
N-FK: Una relación que contiene al menos dos llaves foráneas
PK + S-FK: Una relación que contiene una llave primaria y una llave foránea.
PK + N-FK: Una relación que tiene una llave primaria y dos llaves foráneas.
PK + N-FK: Una relación que tiene una llave primaria y más de dos llaves foráneas.
C-PK + S-FK: Una relación que tiene una llave primaria compuesta y solo una llave foránea
C-PK + N-FK: Una relación tiene una llave primaria compuesta y dos llaves foráneas
C-PK + N-FK: Una relación tiene una llave primaria compuesta y mas de dos llaves foráneas

3.2. Diferentes comandos usados por diseñadores, para producir diseños de bases de datos equivalentes, pero sintácticamente diferentes en el SQL-DDL

Diversos autores están produciendo información semántica diferente basados en un mismo esquema de la BD, causando cierta incertidumbre para los usuarios, sin embargo actualmente los investigadores buscan un consenso para producir siempre todos el mismo resultado, en el trabajo realizado por Sequeda et al [3], son analizados los trabajos realizados por los autores mas emblemáticos considerando principalmente la estructura básica de mapeo en un lenguaje semántico, las formas de interpretar atributos y restricciones, la implementación de la herencia en el lenguaje semántico (LS), la forma en que se interpreta la relación entre las tablas, el lenguaje que es utilizado para almacenar la información semántica generada, y otros aspectos de relevancia.

3.3. La evolución en el alcance del SQL ocasiona un diseño semántico más complejo

Desde la creación de la primer versión de SQL en 1986, el lenguaje ha evolucionado, permitiendo el soporte a bases de datos distribuidas, soporte de XML, manejo recursivo de *triggers* (disparadores), soporte de la tecnología orientada a objetos, además de incrementar la relación semántica entre los datos. Muchos de los analizadores de SQL-DDL no soportan las diferentes versiones existentes entre estos estándares, aún el SQL92 se considera un importante desafío para la automatización semántica, esto debido a la complejidad involucrada en estos comandos, en la actualidad no existe ningún desarrollo que pueda interpretar triggers en la tecnología semántica, esto principalmente debido a la complejidad involucrada en estos comandos. Adicionalmente cada productor de software de bases de datos implementa el

software con pequeñas diferencias, siendo necesario que sea desarrollada una versión por cada motor de base de datos desarrollado.

Tabla 3. Características más importantes de versiones de SQL estándar.

Año	Nombre	Alias	Comentarios
1986	SQL-86	SQL-87	Primera liberación realizada por ANSI.
1992	SQL-92	SQL2, FIPS 127-2	Revision Mayor (ISO 9075), SQL-92.
1999	SQL:1999	SQL3	Se agregan expresiones regulares. Triggers y consultas recursivas. Soporte de aspectos orientados a objetos.
2003	SQL:2003	SQL2003	Se introducen aspectos relacionados con XML Aspectos de secuencias estandarizadas y columnas con valores autogenerados.
2006	SQL:2006	SQL 2006	Son definidas diferentes formas en que SQL puede usarse en conjunción con XML. Importación, almacenamiento y manipulación de XML en una base de datos en SQL. Proporciona aplicaciones para integrar en código SQL el uso de XQuery, el lenguaje de consultas XML publicado por el W3C. Acceso concurrente a datos de SQL y documentos XML.
2008	SQL:2008	SQL 2008	Legaliza la instrucción ORDER BY fuera de los limites de cursores definidos. Incrementa la instrucción INSTEAD OF en triggers. Incrementa el estatuto TRUNCATE .
2012	SQL:2011		En proceso de liberación.

3.4. Lenguajes de ontologías más importantes que permiten expresar diferentes relaciones de datos usando diferentes comandos

Para clarificar las diferencias existentes entre dos de los lenguajes de ontologías más importantes (RDFS y OWL), es importante indicar el rol de cada lenguaje. Primeramente el RDF (Resource Description Framework) define una estructura adicional a las tripletas, las cuales básicamente se conforman de (Sujetos, Predicados y Objetos), el elemento de más alta relevancia en esta representación, es el predicado denominado "rdf:type". El cual se usa para describir que cada atributo de la base de datos, pertenece a un tipo específico de datos.

El lenguaje RDFS (RDF Schema) define clases que representan el concepto de los sujetos, objetos y predicados, esto permite comenzar a crear estatutos acerca de clases de cosas y tipos de una inter-relación. Usando estas clases y tipos es posible establecer relaciones entre dos clases, permite también describir en un lenguaje de texto entendible por humanos el significado de una relación o una clase. Permite igualmente describir los usos legales de varias clases e interrelaciones, sirve igualmente para indicar si una propiedad de clase es un subtipo de un tipo más general. OWL incrementa semántica al esquema, permite especificar más información acerca de las propiedades y las clases, algo importante es que permite expresar la información en tripletas, es posible indicar la transitividad. Permite igualmente el uso de sinónimos, manejo de cardinalidad, rango de datos y muchos comandos más para cono-

cer más específicamente las clases de comportamiento. En la figura 3 se muestra la estructura propuesta por el W3C, puede observarse la relación existente entre RDF, RDFS y OWL y que mientras más se sube de nivel en la figura, los datos producidos son más significativos.

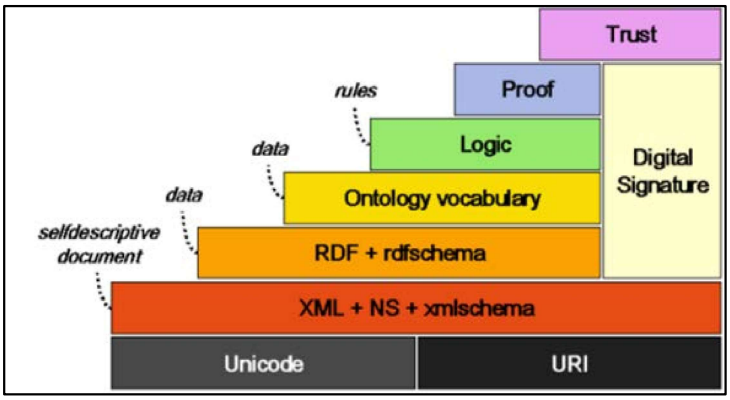

Fig. 3. Niveles en la web semántica propuestos por la W3C.

4. Formulación de ontologías

La definición mas popular de ontología es la propuesta por Gruber: "*Una especificación formal explícita de una conceptualización compartida*" [14]. *Formal* significa que la especificación está codificada en un lenguaje cuyas propiedades formales son bien entendidas; una *especificación explícita* significa que los conceptos y las relaciones en un modelo abstracto proporcionan nombres explícitos y definiciones. Una *Conceptualización* en este contexto, se refiere a un modelo abstracto de como las personas piensan acerca de las cosas, es usualmente restringido a un área particular. *Compartido* se refiere a que el propósito principal de una ontología es para ser usada y reutilizada entre diferentes aplicaciones y comunidades.

Diferentes razones hacen necesaria la creación de una ontología: a) para tener un entendimiento común de la estructura de la información entre personas y agentes de software; b) para habilitar el reuso del dominio de conocimiento; c) para hacer las afirmaciones del dominio explícitas; d) para separar el dominio del conocimiento del conocimiento operacional; e) para analizar el dominio de conocimiento.

Existen muchas discusiones acerca del significado exacto de una ontología, sin embargo existe una coincidencia en las siguientes dos descripciones: a) Es considerado como un vocabulario de términos que se refieren a puntos de interés en un dominio determinado; b) es una especificación acerca del significado de términos basados en algún tipo de lógica. Una ontología junto con algún conjunto de instancias concretas de una clase constituye una base de conocimientos. La taxonomía de clases es un punto fundamental en cada ontología. Es necesario enfatizar que la construcción de una ontología consiste de 6 pasos [20]: 1) Especificación del propósito, uso, alcance y grado de formalidad de la ontología; 2) Colección de datos usando varios métodos de recolección; 3) Conceptualización de términos del dominio (ontología preliminar); 4) Integración con otras ontologías; 5) Formalización en un lenguaje de ontologías; 6)

Evaluación de completitud, consistencia y redundancia. Estos pasos deben ser considerados como base para producir ontologías confiables para el propósito buscado.

4.1. Tipos de ontologías

Existen diferentes formas de clasificación de las ontologías, una posible clasificación se basa en la manera de especificar el significado de los términos. En la figura 4 se muestra en un extremo las ontologías poco formales adecuadas para tareas sencillas, en el extremo contrario se muestran las ontologías extremadamente rigorosas, formalizadas con teorías lógicas formales, estas ontologías proporcionan un soporte para el razonamiento automatizado siendo adecuadas para el descubrimiento de conocimiento.

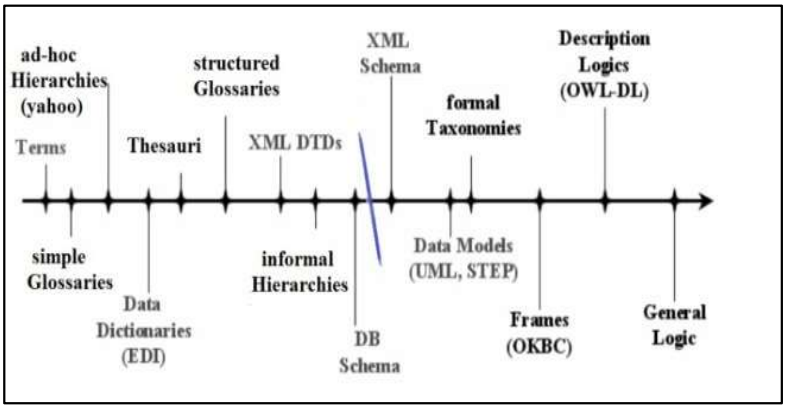

Fig. 4. Clasificación de Ontologías en base a su complejidad

Otra clasificación es con el uso de escenarios de aplicación, siendo de primordial importancia quien interactúa con la ontología producida, en este caso cuatro categorías pueden ser mencionadas de acuerdo a la clasificación realizada por Uschold and Gruninger [15]. Otras posibles clasificaciones de ontologías que pueden ser consideradas son las siguientes: a) De generales a específicas (Genéricas, de Nucleo, de Dominio, de Tareas, y de Aplicación); b) relacionadas con los lenguajes de representación; c) basadas en lógicas de descripción (DL), las cuales pueden ser usadas para representar terminología de conocimiento de una aplicación de dominio en una forma formal y estructurada; d) Caracterización de Ontologías DL en base al constructor utilizado. Las ontologías pueden utilizar varias de estas clasificaciones al mismo tiempo, el incremento de constructores en una ontología incrementa la complejidad de una ontología producida, es por esto que es necesario antes de crear una ontología un buen conocimiento del problema a resolver Kepler et al [16].

4.2. Ontologías y bases de datos

En vista de que los esquemas de las BD son documentos formales diseñados por especialistas en ingeniería de software que buscan modelar el mundo real, es posible producir una ontología a partir de éstos, las ontologías están formadas por una se-

cuencia de anotaciones, axiomas y hechos, los elementos más esenciales son los axiomas y los hechos, los cuales proporcionan información acerca de las clases, propiedades e individuos, de acuerdo a Kashyap [19], la construcción de ontologías a partir de BD relacionales requiere analizar el esquema y determinar Llaves primarias y foráneas y dependencias de inclusión. Se considera que el lenguaje OWL-LITE es suficientemente poderoso para representar la información almacenada en el esquema, aspectos tales como: Igualdad de términos, características de propiedad, restricciones de propiedad, restricciones de cardinalidad, información de encabezados, intersección de clases, control de versiones, propiedades de anotación y tipos de datos son elementos expresables en el lenguaje para crear una ontología de una base de datos de intranet.

5. Generación semiautomática de ontologías

Como se mencionó en la sección 2, la conversión automática de información contenida en las bases de datos relacionales, se encuentra todavía bajo un proceso de refinamiento, es por esto que la mayor parte de desarrolladores de ontologías utilizan la generación semiautomática de ontologías, proceso que puede conformarse por tres pasos para preparación de la información y al menos otro paso para la explotación de la información semántica producida, a continuación se mencionan los pasos requeridos para generar un ontología de manera semiautomática

Conversión de la estructura de bases de datos relacionales

Archivos RDF (conversión física de datos). Muchos productos de software de tecnología semántica, han adoptado este método, en el cual el total de los datos son convertidas a tripletas de información (s,p,o), una vez realizada la conversión, se incluye información acerca de la estructura de los datos, usando generalmente el lenguaje RDFS, y procediendo posteriormente a alinear la información producida a una ontología de dominio existente (wrapping systems), algunos ejemplos de este enfoque pueden encontrarse en : Arens et al 1993 [4], Barraza et al 2006 [5] y Zhou et al [6].

Archivos Map (Conversión de la estructura de datos). En este enfoque el esquema de SQL es transformado a una ontología sencilla, pudiéndose tener 2 enfoques diferentes, el primero sería que los datos se dejan en las bases de datos y se establecen mecanismos para mapear los datos relacionales a semánticos, y el segundo es que los datos relacionales son generados como instancias de la ontología creada, ejemplos del primer caso se pueden encontrar en D2RQ, y del segundo caso en los autores comparados en el trabajo realizado por Sequeda et al [3] y por autores como Shen et al.

Generación de ontologías

Generación manual de ontologías. El proceso de crear ontologías, es generalmente realizado utilizando editores de ontologías, los cuales permiten organizar gráficamente los conceptos componentes de la ontología, existen muchas herramientas para realizar ésta actividad, dentro de las principales podemos mencionar a Protegé, Swoop, Topbraid, etc. Estas herramientas nos ayudan a generar las partes repetitivas

de la ontología, evitando la generación de errores debido a fallas en la sintaxis, permitiendo visualizar la ontología en el momento en que se está creando.

Generación automática de ontologías. Este tópico requiere tener un completo dominio del estándar SQL-DDL, el cual permite generar la asociación existente entre los datos almacenados y la base de datos en términos semánticos, en este enfoque existen diversos desajustes, debido a que diferentes autores interpretan los comandos SQL de manera diferente, o utilizan diferentes comandos para generar las ontologías. Actualmente se busca la estandarización en esta área para producir exactamente las mismas ontologías

Interconexión de datos semánticos con ontologías

En la actualidad este aspecto es realizado manual o semiautomáticamente, produciendo una salida de los datos relacionales y creando una asociación con la ontología utilizando un editor de ontologías , ésta asociación requiere la consideración de similaridades, como se menciona en los trabajos de Buttler [8] y de Todorov y Geibel [9], donde se consideran los siguientes aspectos: a) Similaridad en strings, d) Similaridad en sinónimos, c) Similaridad basada en instancias. Para la alineación de términos, se ha considerado la utilización de dos tipos de similitud (a y b) debido a que el tipo de aplicación a integrar, está más relacionada con la estructura de los datos y no tanto con el contenido de los atributos En el trabajo propuesto se considera la utilización de 2 tipos de similitud, más relacionados con la estructura de los datos que con las instancias.

Resolución de consultas utilizando la tecnología semántica

Esta parte del trabajo, se encuentra actualmente en desarrollo, se están considerando dos diferentes opciones: a) Utilizar el lenguaje de consultas semánticas en archivos RDF (SPARQL), b) Desarrollar un lenguaje de consultas semánticas aplicado a bases de datos relacionales.

6. Areas de investigación futuras

Existen en la actualidad varios problemas abiertos, los cuales pueden categorizarse en tres grandes grupos, tales como:

- **Mapeo de Esquemas:** En este proceso, no obstante el gran número de productos existentes, se requiere realizar algunos refinamientos, debido a que no existen suficientes mecanismos para mapear la información de las bases de datos, tampoco existe soporte para todos los manejadores de BD, sin embargo existe soporte para la mayor parte de los productos más conocidos. El lenguaje para realizar ontologías más actual es el OWL, el cual conforma las ontologías con una secuencia de anotaciones, axiomas y hechos. Los elementos mas relevantes son los axiomas y los hechos, ya que proporcionan información acerca de las clases, propiedades e individuos, importantes aspectos a considerar para producir una ontología exitosa, son detallados por Du Bois et al [20].

- **Razonamiento con búsquedas imprecisas:** Una de las fuerzas de la utilización de ontologías, es la relacionada con la asociación de términos en una jerarquía dentro de un dominio de conocimiento, esto permite el uso de sinónimos y búsquedas más generalizadas, proporcionando una base solida para el conocimiento estructurado, con la infraestructura semántica es posible manejar términos equivalentes, así mismo permite realizar actividades de agregación o desagregación como se necesite [22].
- **Consultas Semánticas:** Una comparación entre los lenguajes de consultas de las BD y los lenguajes de consulta de ontologías, es necesaria para seleccionar la mejor estrategia para la realización de consultas semánticas. Dos diferentes estrategias pueden ser utilizadas en los lenguajes de consulta ontológicos: 1) Uso exclusivo de clases y propiedades, los cuales se utilizan para capturar relaciones intencionales, siendo pobres como lenguajes de consultas, debido a que no es posible referirse al mismo objeto por diferentes trayectorias de navegación dentro de la ontología. 2) Full SQL (equivalente a lógica de primer orden), tiene el inconveniente de ser indecidible cuando se tiene información incompleta Giacomo et al [23], proponen el uso de uniones de consultas conjuntivas (CQs).

7. Conclusiones

La extracción de información almacenada en bases de datos usando ingeniería inversa para su conversión a información semántica, requiere que la BD esté en al menos en 3NF (tercer forma normal), debido a que esto ayuda a asegurar que el diseño está adecuadamente realizado, este proceso involucra que una vez que se tiene toda la información automatizada, se cuenta con un uso adecuado de llaves primarias y una correcta asociación entre tablas. Sin embargo un procesamiento adicional debe ser hecho para contar con una ontología funcional, es necesario trabajar con los pasos recomendados por Du Bois [20] para crear las ontologías, lo cual no se considera una tarea sencilla. La elaboración de ontologías requiere parsear el script de SQL y comenzar identificando los elementos relevantes necesarios para crear la ontología. El grupo incubador W3C [12] continua la investigación a través del grupo RDB2RDF para mejorar el proceso de conversión automática de datos, la integración de información de diversos repositorios, optimización de almacenamiento y optimización de consultas. Es importante mencionar que de las dos alternativas para producir ontologías, consideramos el "mapeo directo" como la mejor alternativa a ser realizada, aunque no es el mas popular debido a su complejidad, es importante mencionar que la generación automática aún se encuentra bajo investigación y desarrollo [10], una estandarización es fundamental en el área, la cual todavía no se ha consolidado, aun cuando las ontologías sean producidas manualmente, estas no están totalmente aceptadas por la comunidad de productores de ontologías, como puede ser visto en la investigación de Todorov and Geibel [9]. La creación automática de ontologías no se encuentra todavía estandarizada, existen diferentes iniciativas para mejorar el soporte y desempeño de consultas cuando existe información incompleta, una de las características para el logro de la integración semántica es el incremento de posibilidades para la explotación de consultas así como una simplificación en el proceso de explotación.

Referencias

1. Jung An, Geller J, Wu Y, Choon S.: Semantic Deep Web: Automatic Attribute Extraction from the Deep Web, Data Sources, ACM 1-59593-480 (2007)
2. Berners Lee, Tim; Fischetti, Mark: Weaving the web. Harper San Francisco, chapter12, ISBN 978-0-06-251587-2 (1999)
3. Sequeda J., Tirmizi S-, Corcho O., Miranker D.: Survey of directly Mapping SQL databases to the Semantic Web. The Knowledge Engineering Review, Cambridge University Press, Vo 26:4, 445-486 (2011)
4. Arens, A. Chee, C. Y., Hsu, C & Knoblock, C.: Retrieving and integrating data from multiple information sources. International Journal on Intelligent and Cooperative Information Systems, 127-158, (1993)
5. Barrasa, J., Gómez Pérez, A.: Upgrading relational legacy data to the semantic web. In: proc. of 15th international conference on World Wide Web Conference (WWW 2006), pages 1069,1070, Edinburgh, United Kingdom (2006)
6. Zhou Shufeng: Mapping Relation Databases for Semantic Web. International Conference on Future BioMedical Information Engineering (2009)
7. Shen, G., Huang, Z., Zhu, X. & Zhao, X: Research on the rules of mapping from relational model to OWL. In: Proceedings of the workshop on OWL: Experiences and Directions, Athens, GA, USA (2006)
8. D. Buttler: A short Survey of Document Structure Similarity Algorithms. International Conference on Internet Computing, Las Vegas, NV, United States (2004)
9. K. Todorov, P Geibel: Ontology Mapping via structural and Instance-Based Similarity Measures. Third International Workshop On Ontology matching, OM2008 (2008)
10. D O´Leary: Different Firms, Different Ontologies, and No one Best Ontology. IEEE Intelligent Systems, IEEE, pp 72-78 (2000)
11. W3C Semantic Web Activity, http://www.w3.org/2001/sw/
12. W3C RDB2RDF Incubator Group 2009, http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf
13. A. Doan and A. Halevy: Semantic-Integration Research in the Database Community. AI Magazine Volume 26 Number 1 (2005)
14. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5:199-220 (1993)
15. M. Uschold, M. Gruninger: Ontologies and Semantics for Seamless Connectivity, SIGMOD Record, Vol. 33, No. 4 (2004)
16. F. Kepler, C. Paz-Trillo, J. Riani, M. Moretto , K. Valdivia-Delgado, L. Nunes, and R. Wassermann: Classifying ontologies. In Proceedings of the Second Workshop on Ontologies and their Applications (WONTO 2006) (2006)
17. OWL Web Ontology Language Overview, W3C Recommendation 10 February (2004)
18. Asunción Gómez-Pérez and Oscar Corcho: Ontology Languages for the Semantic Web. IEEE INTELLIGENT SYSTEMS (2002)
19. Kashyap, V.: Design and creation of ontologies for environmental information retrieval. 12th Workshop on Knowledge Acquisition Modeling and Management (KAW´99). Banff, Canada, October 1999, http://sern.ucalgary.ca/ksi/kaw/KAW99/papers/Kashyap1/kashyap.pdf (2007-03-15).
20. Du Bois, B.: Towards an ontology of factors influencing reverse engineering. In: STEP '05: Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice, USA, pp. 74–80 (2005)
21. Bizer C., Cyganiak, R.: D2RQ — Lessons Learned. Position paper for the W3C Workshop on RDF Access to Relational Databases, Cambridge, USA (2007)

22. Davidson, James y Jerrold Kaplan: Natural Language Access to Data Bases: Interpreting Update Requests. Computer Science Department, Stanford University, California, ACM Press (1983)
23. G de Giacomo, Towards Systems for Ontology-based Data Access and Integration using Relational Technology, Sapienza Universita di Roma, U of Toronto (2010)