Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Tecnologías de bases de datos
Recibido: 9/10/2013 | Aceptado: 20/11/2013 | Publicado: 21/01/2014

Búsqueda de correspondencia entre esquemas conceptuales de datos

Searching of correspondences between data conceptual schemata

Javier Agustín González[1*], Frank Reyes García [2], Abel Rodríguez Morffi [2], Luisa M. González González [2]

[1] Centro de Investigación en Bioalimentos. Carretera Patria, km 1 ½, Morón, Ciego de Ávila. Cuba
[2] Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba

*Autor para la correspondencia: javier@cibcav.cu
{frank26; arm; luisagon}@uclv.edu.cu

Resumen

En la modelación conceptual de datos, los usuarios y diseñadores pueden modelar diferentes vistas de un dominio en la que los requerimientos se formalizan en varios esquemas conceptuales; los esquemas así obtenidos usualmente presentan una heterogeneidad tanto semántica como estructural. La herramienta MERMAID ayuda en la modelación de dominios y utiliza el formato MXP para almacenar los esquemas resultantes. Existen herramientas que realizan la búsqueda de correspondencia entre esquemas conceptuales tales como S-Match, Cupid y COMA. El formato MXP no es válido para ninguna de estas aplicaciones; por lo que se propone la creación de un operador de correlación calculado con diferentes combinaciones de métricas, apoyado en relaciones semánticas y con un valor de similitud sintáctica. Se consideró la selección y adaptación de técnicas existentes de descubrimiento de correspondencia, la combinación de resultados de diferentes técnicas sintácticas y la clasificación de las correspondencias en tres alternativas. El operador clasifica las relaciones semánticas en equivalencia, desigualdad, más general y menos

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

general. El valor de similitud sintáctica solo se ofrece para las relaciones de equivalencia ya que en las demás carece de sentido. La clasificación de las correspondencias brindadas por el operador se optimizan agregando nuevas técnicas de similitud sintáctica e integrando un tratamiento de la incertidumbre.
Palabras clave: Correspondencia semántica, correspondencia sintáctica, similitud.

Abstract

In the context of conceptual data modeling, view integration refers to the activity of integrating and unifying different conceptual schemata modeled over a universe of discourse in a global schema. The integration process includes complex tasks such as identifying common concepts between views, determining appropriate structures and discovering inter-schemes properties. Searching correspondences between conceptual schemata is a critical and non trivial task that usually is done manually, which obviously has major limitations. In the last 15 years many researchers have dedicated efforts to discover and combine techniques in the endeavor for automating the process of discovering correspondences between schemata. Only partial solutions to specific domain applications have been proposed. This paper proposes a match operator for data conceptual schemata based on the combination of syntactic and semantic match operators.
Keywords: *Correspondences, schema matching, semantic matching, similarity, syntactic matching.*

Introducción

En la modelación conceptual de datos, la integración de vistas se refiere a la actividad de integrar y unificar en un esquema global los diferentes esquemas conceptuales modelados sobre un mismo universo de discurso. Como consecuencia, los elementos coincidentes en las distintas vistas pueden estar representados de manera heterogénea, tanto estructural como semánticamente. La heterogeneidad entre los elementos comunes de las vistas a integrar constituyen fuentes de conflictos que deben ser resueltos para garantizar que el esquema integrado o esquema global sea consistente. La búsqueda de correspondencias (BC) entre esquemas conceptuales de datos es una tarea crítica y no trivial que usualmente se realiza manualmente. En los últimos años diferentes investigadores han dedicado esfuerzos al descubrimiento y combinación de técnicas en el empeño de automatizar el proceso de descubrimiento de correspondencias entre esquemas; algunos de estos trabajos son (Doan and Madhavan, 2002; Rahm and Bernstein, 2001; Bernstein and Madhavan, 2001; Giunchiglia and Shvaiko, 2004; Gal, 2006; Hong and Rahm; Berlin and Motro, 2001; Madhavan and Bernstein, 2001; Shvaiko and Euzenat, 2005; Dong and Halevy, 2007; Gal and Martínez, 2009; Cheng and Gong, 2010; Miller and Hernández, 2001; Melnik and Rahm, 2003), hasta ahora solo se han propuesto soluciones parciales para dominios de aplicación específicos como COMA (Hong and Rahm), *Cupid* (Madhavan and

Bernstein, 2001) y *Similarity Flooding* (Melnik and Rahm, 2003). En este artículo se plantea la selección y adaptación de técnicas existentes de descubrimiento de correspondencias entre esquemas al dominio de esquemas conceptuales de datos. En la solución propuesta se combinan resultados de diferentes operadores de correspondencia sintáctica y semántica.

Materiales y métodos

Búsqueda de correspondencia

La BC entre esquemas es el proceso mediante el cual se identifican las posibles correspondencias de similitud que se pueden establecer entre elementos de diferentes esquemas. Una correspondencia es una relación entre uno o más elementos de un esquema y uno o más elementos de otro esquema. Una de las primeras contribuciones expuestas en (Doan and Madhavan, 2002) fue que las técnicas de correspondencias desarrolladas por separado como parte de aplicaciones se podrían combinar en un árbol de conocimiento y los resultados de estas serían utilizados en la correspondencia de esquemas como un objetivo separado e independiente de la aplicación que las use. De esta manera se comenzó a trabajar en la correspondencia como un tema independiente sobre el cual (Bernstein and Madhavan, 2011) expone una taxonomía. En (Madhavan and Bernstein, 2001) se resumen las características de los operadores de correspondencia basado en múltiples algoritmos. Esto permite seleccionar los métodos para la BC dependiendo de la aplicación de dominio y el tipo de esquema. La BC a nivel de esquema solo considera la información del esquema pero no de las instancias; la información que se presenta son las propiedades de los elementos a considerar lo cual incluye su etiqueta, el nombre de los atributos, el tipo de dato, la descripción, el tipo de relación, las estructuras y las restricciones. Para cada elemento de un esquema candidato a establecer una correspondencia con algún elemento de otro esquema candidato, la correspondencia estima el grado de similitud por un valor numérico normalizado entre 0 y1, con posibilidades de identificar los mejores candidatos a corresponder, como se explica en (Doan and Madhavan, 2002). En la granularidad de las correspondencias se distinguen dos alternativas: correspondencias a nivel de elemento y a nivel estructural. Dado dos esquemas como entrada, para cada elemento del primer esquema se determinan los elementos correspondientes en el segundo esquema. En el caso más simple, solo se consideran elementos al nivel más fino de granularidad; a este nivel se le llama nivel atómico, como son los atributos en un esquema.

Operador de correspondencia

En la definición de un operador de correspondencia se necesita escoger una representación para los esquemas de

entrada y para el conjunto de correspondencias de salida, aparte de explorar muchas aproximaciones de correspondencias basadas en diferentes criterios. Este operador depende de un conjunto de tipos de informaciones contenidas en los esquemas que debe extraer, organizar según sus necesidades y ser capaz de interpretar los resultados de los algoritmos para poder combinar los resultados. Así, para el propósito de este trabajo, se define el esquema como un conjunto de elementos conectados por alguna estructura, que puede ser un modelo orientado a objeto, un documento XML o un grafo dirigido; la estructura seleccionada es un grafo dirigido. Los elementos del grafo son los vértices que representan los conceptos y las aristas que representan las relaciones entre los conceptos; por lo que el centro de la atención es la información contenida en los elementos (vértices y aristas) ya sean etiquetas o conceptos.

Clasificación de correspondencias usadas

Las correspondencias usadas en la búsqueda son las aproximaciones lingüísticas y semánticas. De acuerdo con lo expuesto en (Doan and Madhavan, 2002), las correspondencias lingüísticas o basadas en lenguaje usan nombres o textos (por ejemplo, palabras o sentido de las palabras) para encontrar similitud semántica entre los elementos de los esquemas. Esta correspondencia puede ser definida de varias formas, ya sea como igualdad de nombres o como igualdad de representaciones después de un procesamiento; esto es importante porque se trabaja con sufijos, prefijos y símbolos especiales. La similitud de nombres en (Bernstein and Rahm, 2001) describe técnicas basadas en búsqueda de sub-cadenas comunes, distancias de edición, fonética, cantidad de caracteres, alineaciones de cadenas y las correspondencias provistas por el usuario. La aproximación semántica explora relaciones semánticas entre palabras que requieren del uso de diccionarios, tesauros y taxonomías que contienen relaciones tales como sinonimia, holonimia, hiponimia, hiperonimia, meronimia y antonimia.

Las relaciones semánticas son definidas en (Giunchiglia and Shvaiko, 2004) como: relaciones equivalentes, más generales, menos generales y desiguales. Existen dos niveles de granularidad en el desempeño de la correspondencia tanto semántica como sintáctica. A nivel de los elementos las técnicas de búsquedas calculan las correspondencias de elementos entre las etiquetas individuales y conceptos en los nodos; a nivel estructural las técnicas calculan la correspondencia de los elementos entre los subgrafos. Las técnicas de correspondencia semántica a nivel de elemento analizan etiquetas individuales o conceptos en los nodos. En este nivel se pueden explorar todas las técnicas descritas en la literatura, véase por ejemplo (Bernstein and Rahm, 2001; Melnik and García, 2002). La principal diferencia aquí es que, en lugar de una medida de similitud sintáctica, estas técnicas se deben modificar para devolver una relación semántica R, tal como las que se definieron anteriormente. Las técnicas de semántica débil son las técnicas de

manejo de sintaxis; ejemplo de estas técnicas son las que consideran las etiquetas de los nodos como cadenas, o las que analizan un tipo de dato, o algunas basadas en fonética como soundex que genera un código para cada cadena. Las técnicas de semántica fuerte exploran, en el nivel de elemento, la semántica de las etiquetas. Estas técnicas se basan en el uso de herramientas que codifican explícitamente la información semántica, las ontologías, la base de datos léxica WordNet o combinaciones de ellos. Nótese que estas técnicas se utilizan también en concordancia sintáctica. En este último caso, sin embargo, la información semántica se ha perdido antes de pasar al análisis de correspondencias a nivel estructural y codifican aproximadamente a las relaciones sintácticas.

WordNet

WordNet es un recurso lingüístico ideado para el uso automático, el cual posee información psicolingüística y está organizado en base a los significados de las palabras basado en la estructura de tesauros. Este recurso contiene todos los aportes que manejan los diccionarios electrónicos y los diccionarios tradicionales, además contiene descripciones basadas en conceptos y relaciones psicolingüísticas entre las palabras según se expresa en (Troyano). Los synsets constituyen la estructura básica de la base léxica. En él se guarda el significado, las formas que presentan una relación de sinonimia en dicho significado y pequeñas frases que brindan más información. Además el synset tiene un identificador numérico único y una serie de relaciones con otros synsets las cuales representan las relaciones semánticas.

De acuerdo con lo que se expresa en (Giunchiglia and Shvaiko, 2004), se tienen los siguientes casos:

- equivalencia: un concepto es equivalente a otro si y solo si existe al menos un sentido del primer concepto el cual es sinónimo de un sentido del segundo concepto;
- más general: un concepto es más general que otro si al menos existe un sentido del primer concepto el cual tiene un relación de hiponimia o meronimia con el segundo concepto;
- menos general: un concepto es menos general que otro si y solo si existe al menos un sentido del primer concepto que tiene un sentido del segundo concepto como un hipérnimo o como un holónimo;
- la desigualdad: dos conceptos son desiguales si ellos tienen dos sentidos (uno de cada uno) los cuales son diferentes hipónimos de un mismo synset o si ellos son antónimos.

Incertidumbre en la correspondencia

La investigación sobre correspondencias entre esquemas ha sido desarrollada por más de 25 años, primero como parte del proceso de integración y luego como parte de un campo de investigación autónomo, véanse los artículos (Batini

and Lenzerini, 1986; Doan and Madhavan, 2002; Sheth and Larson, 1990; Shvaiko and Euzenat, 2005). Durante años, un grupo significativo de trabajos se consagraron a la identificación de correspondencias de esquemas y heurísticas para dichos correspondencias.

El objetivo principal de este proceso es proporcionar correspondencias que sean eficaces desde el punto de vista del usuario, manteniendo un orden computacional aceptable. Ejemplos de estas herramientas algorítmicas incluyen a COMA (Hong), *Cupid* (Madhavan and Bernstein, 2001), *Onto-Builder* (Gal and Modica, 2005), Autoplex (Berlin and Motro, 2001), *Similarity Flooding* (Melnik and Rahm, 2003), *Clio* (Miller and Hernández, 2001), Glue (Doan and Madhavan, 2002), y otros (Bergamaschi and Castano, 2001; Castano and Antonellis, 2001; Saleem and Bellahsene, 2007). Éstos han surgido de diferentes comunidades científicas, incluyendo administración de base de datos, recuperación de información, ciencias de la información, semántica de los datos, Web semántica, entre otros. Los documentos de investigación de las diferentes comunidades muestran que existen aristas de solapamiento, similitud y en algunos se evidencian resultados equivalentes.

A través de los años, la BC entre esquemas ha emergido con una incertidumbre heredada; ya que se pueden considerar múltiples correspondencias posibles y al elegir se puede tomar una decisión errónea como se explica en (Gal, 2006). Un motivo fundamental de la incertidumbre de este proceso es la enorme ambigüedad y heterogeneidad de los conceptos de descripción de datos. En (Miller and Haas, 2000) se justifica la incertidumbre sobre la base de que "la representación sintáctica de esquemas y datos no transmiten completamente la semántica de las diferentes bases de datos"; es decir, la descripción de un concepto en un esquema puede ser semánticamente engañosa. Desde el 2003, el trabajo sobre la incertidumbre en la correspondencia de esquema se ha recuperado con resultados como los reportados en (Dong and Halevy, 2007; Gal and Martinez, 2009; Magnani and Rizopoulos, 2005; Cheng and Gong, 2010).

Combinación de correspondencias

Tanto en (Bernstein and Rahm, 2001) como en (Bernstein and Madhavan, 2011) se han revisado diferentes variantes de correspondencias; cada cual usa un tipo útil de información. Por eso, es poco probable que un solo algoritmo de BC obtenga buenos valores de similitud entre los elementos de los esquemas. Esto puede ser mejorado con los algoritmos híbridos que integran múltiples tipos de correspondencias. El análisis a nivel estructural también se beneficia con el uso de otros criterios. Así se puede usar un algoritmo para generar una parte de las correspondencias y otros criterios para completarlas. Por otro lado, se puede usar una correspondencia compuesta, que combina los

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

resultados de diferentes correspondencias individuales incluyendo a las correspondencias híbridas; esta habilidad de combinar correspondencias es más flexible que el de los híbridos. Una correspondencia híbrida usa una combinación de técnicas de correspondencias particulares que se ejecutan simultáneamente o en un orden fijo. Una correspondencia compuesta permite seleccionar de un repertorio modular, por ejemplo sobre una aplicación de dominio. Es más, una correspondencia compuesta debe permitir una clasificación flexible de correspondencia para que ellos se ejecuten simultáneamente o secuencialmente. En el último caso, el resultado de la primera correspondencia es consumido y extendido por una segunda correspondencia para lograr una mejora reiterativa del resultado inicial.

En la combinación de los resultados del cálculo de la similitud usando los diferentes algoritmos escogidos por el usuario al realizar la BC se utilizó el procedimiento que se propone en (Hong and Rahm). La combinación se realiza en tres etapas principales; el llenado del cubo de similitud, la obtención de la matriz de similitud y el cómputo de los resultados. Tal como se muestra en las Figuras 1 y 2 traducidas de (Hong and Rahm).

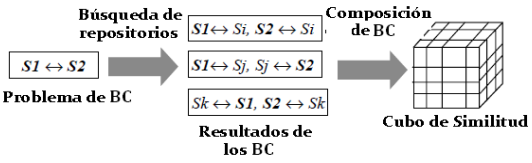

Figura 1. Obtención del cubo de similitud.

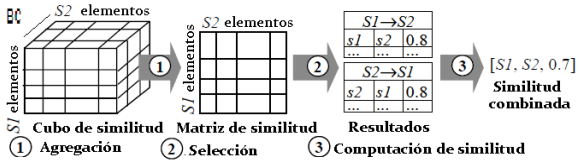

Figura 2. Proceso de obtención de similitud.

Los dos primeros pasos son obligatorios, mientras que el tercero es opcional. El primer paso se conoce como "Agregación de resultados de correspondencias específicas", el segundo se llama "Selección de candidatos" y el tercero es "Cálculo de similitud combinada". En los dos primeros pasos se realiza la combinación de los valores de similitud para obtener un resultado completo de la correspondencia. El paso tres se usa para las correspondencias híbridas. El cubo de similitud está formado por los elementos del primer esquema en las filas y por los elementos del segundo esquema en las columnas. Las dimensiones del cubo forman los diferentes algoritmos que se van a aplicar en el proceso, de manera tal que si se usan en un proceso n algoritmos el cubo tendrá n dimensiones. Luego de calcularse todos los valores de similitud para cada combinación de elementos en el cubo, se realiza el primer subpaso. Los valores del cubo son agregados a un valor de similitud combinada para cada par de elementos de los elementos esquemas. Con m elementos en S1 y n elementos en S2 se obtiene una matriz de m x n de valores de similitud combinada. Este subproceso usa tres tipos de estrategias de agregación:

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

- Máximo: escoge al máximo valor de similitud de cualquier correspondencia; es una estrategia optimista tomando lo máximo complemento de cada correspondencia.
- Promedio: determina el promedio de los valores de similitud; es una estrategia más realista.
- Mínimo: se escoge el menor de los valores de similitud; es totalmente opuesta a máximo; es una estrategia pesimista.

La matriz de similitud de los resultados de la operación de agregación sobre las correspondencias tiene la característica de ser pesimista, optimista o estar entre las dos siguiendo el promedio. El conjunto de datos que se representa en la matriz es el usado para complementar las relaciones semánticas construidas. El proceso de filtrado elimina todas las relaciones que tengan un valor de similitud menor que un umbral dado u obtenido de acuerdo con los requerimientos del usuario.

Operador JMatcher

El operador propuesto en este artículo tiene como entrada dos modelos conceptuales. Así el problema de la BC heterogénea se divide en dos pasos:

1. Extraer grafos de los modelos conceptuales.
2. Buscar correspondencias entre los grafos resultantes.

Al definir la noción de una correspondencia entre grafos de forma más precisa, se tiene que un elemento de correspondencia es una cuádrupla <mID, Ni1, Nj2, R>, i=1...h; j=1..k; donde mID es un identificador único del elemento de correspondencia dado; Ni1 es el i-ésimo nodo del primer grafo, h es el número de nodos en el primer grafo; Nj2 es el j-ésimo nodo del segundo grafo, k es el número de nodos en el segundo grafo; y R especifica una relación de similitud de los nodos dados. Una correspondencia es un conjunto de elementos que se corresponden.
La BC entre esquemas permite descubrir correspondencias entre los dos grafos y se clasifica en semántica y sintáctica dependiendo de cómo se calculan los elementos de la correspondencia y de qué tipo de relación de similitud R se use. En la correspondencia sintáctica, la intuición clave es el hacer corresponder a las etiquetas de los nodos, y para buscar la similitud se usan técnicas basadas en el manejo de sintaxis y las métricas de similitud sintáctica. Por lo tanto, en el caso de coincidencia sintáctica, los elementos de correspondencia se calculan como cuádrupla <MID, Li1, Lj2, R>, donde Li1 es la etiqueta en el nodo i del primer esquema, Lj2 es la etiqueta en el j-ésimo nodo del segundo esquema; y R especifica una relación de similitud en la forma de un coeficiente, que mide la similitud entre las etiquetas de los

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

nodos propuestos. Ejemplos típicos de relación de similitud R son los coeficientes en el rango [0,1], como es el caso de los coeficientes de similitud. Los coeficientes de similitud suelen medir la proximidad entre los dos elementos, ya sea de forma lingüística o estructural. Por ejemplo, con base en el análisis lingüístico, el coeficiente de similitud entre los elementos de "teclado" y "tecla" de los dos esquemas hipotéticos podría ser de 0,7. A partir de su nombre, en la correspondencia semántica, la intuición clave está en el mapa de los significados (conceptos). Así, en el caso de coincidencia semántica, los elementos de correspondencia se calculan como 4-tuplas <MID, Ci1, Cj2, R>, donde Ci1 es el concepto del nodo i-ésimo del primer grafo; Cj2 es el concepto del j-ésimo nodo del segundo grafo; y R especifica una relación de similitud en la forma de una relación semántica entre las extensiones de los conceptos en los nodos propuestos. Los posibles valores de R son las relaciones semánticas descritas anteriormente. La figura 3 muestra los dos tipos de correspondencias, y el tipo de resultado ya sea un número o una relación.

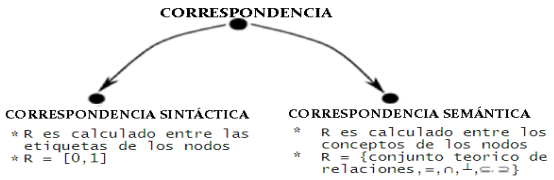

Figura 3. Tipos de correspondencias.

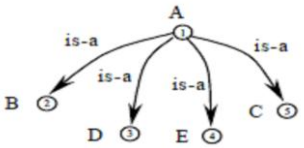

Figura 4. Ejemplo de árbol jerárquico.

Una de las principales diferencias entre correspondencia sintáctica y semántica es que, en la correspondencia sintáctica, cuando se comparan dos nodos, el significado que (implícitamente) se asocia a ellos solo depende de sus etiquetas, independientemente de su posición en el grafo. En la correspondencia semántica, en cambio, cuando coinciden dos nodos, los conceptos que se analizan dependen no solo en el concepto asociado al nodo (el concepto denotado por la etiqueta del nodo), sino también de la posición del nodo en el gráfico. Considérese el ejemplo de la Figura 4. Los números en los círculos son los identificadores únicos de los nodos en cuestión. En la Figura 4, A representa la etiqueta de un nodo, CA representa el concepto denotado por A, Ci representa el concepto en el nodo i. En lo que sigue, a veces se confunden los conceptos con sus extensiones. Por ejemplo, el análisis llevado a cabo cuando el nodo numerado 5 se somete a una correspondencia (contra un nodo en otro grafo); se intenta hallar correspondencia con la etiqueta del nodo 5, nombrada C. En la correspondencia semántica, en cambio, el comparador intenta hacer coincidir el concepto en el nodo 5, a saber, C5, el cual es un subconjunto de la extensión de CA que son también las extensiones de CC. Así, $C5 = CA \cap CC$. Un comparador semántico por lo tanto, tratará de hacer

corresponder CA ∩ CC y no C. En (Madhavan and Bernstein, 2001) se describe Cupid, algoritmo para encontrar correspondencias sintácticas y en (Giunchiglia and Shvaiko, 2004) se detalla S-Match, algoritmo para encontrar correspondencias semánticas. Ambos están dirigidos a encontrar similitudes en estructuras arbóreas, por lo que no pueden ser aplicados a esquemas conceptuales como ER o a los grafos de dependencia de existencia de MERODE, metodología orientada por modelos para el desarrollo de software, descritos en (Snoeck, 2011). Para la metodología MERODE se creó la herramienta MERMAID a la cual se le adiciona un módulo de ayuda a la integración de esquemas conceptuales de datos que incluye el operador propuesto en este trabajo, con el propósito de ofrecer una primera sugerencia a los usuarios sobre las correspondencias entre los esquemas a integrar. El flujo de trabajo del sistema está dado por el modo de operar de un operador de correspondencia. En la figura 5 se presenta la arquitectura del operador. En la primera fase se realiza la auditoria de datos; es decir, luego de extraer los grafos de los esquemas, se extraen las etiquetas de cada elemento de los grafos, ya sea una arista o un nodo. Después, cada etiqueta se separa en palabras y cada una de estas constituye un token; de estos se expanden aquellos que sean acrónimos o que no tengan significado definido y finalmente se verifica que existan en la base de datos léxica. El proceso de división de etiquetas en palabras consiste en tomar los nombres de los elementos de los grafos y dividirlos en palabras como se realiza en (Madhavan and Bernstein, 2001). Para esto se toman como delimitadores de cadenas cualquier carácter que no pertenezca al alfabeto. En la tabla 1 se muestra un conjunto de nombres de elementos de un esquema con su respectivo conjunto de palabras.

Tabla 1. División de etiquetas.

Nombre de los elementos	Lista de etiquetas obtenidas
Show_Offert	Show, Offert
AlternativeOffert	Alternative, Offert
Num%Id	Num, Id
Address.thesther	Address, theather

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

Figura 5 Arquitectura del operador.

En una verificación prematura, para cada grafo se obtienen todas las palabras clasificadas (como sustantivos, adjetivos, adverbios o verbos) que están implícitas en la representación de los conceptos y se almacenan en un diccionario, tomando como llave del diccionario la palabra y como contenido del mismo a los sentidos que puede tener la palabra en diferentes contextos; para eso se usa WordNet, como se manifestó anteriormente. Por otro lado, el proceso de expansión de las cadenas tiene como objetivo tomar todos los tokens del proceso anterior que WordNet no pudo clasificar y utiliza la retroalimentación proporcionada por el usuario para obtener, a partir de estas, palabras que se puedan clasificar. En la etapa final de verificación se toma cada una de las palabras retornadas por la expansión realizada por el usuario. Aquí se utiliza WordNet para clasificar las nuevas palabras en sustantivo, adjetivo, adverbio o verbos.

En la segunda fase se buscan las relaciones semánticas entre los tokens de ambos grafos basado en el significado de las palabras usadas para nombrarlos. Primeramente se construyen los conceptos simples o complejos según la cantidad de palabras utilizadas para nombrar cada elemento del grafo. Luego se buscan las relaciones semánticas entre dichos conceptos. El algoritmo que se implementó para la búsqueda es una variante del método SMatch expuesto en (Giunchiglia and Shvaiko, 2004) que busca correspondencias semánticas; a éste se incluyen todas las características de SMatch y se adicionan métricas lingüísticas para la búsqueda de correspondencia sintáctica (véase la tabla 2).

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

Tabla 2. Métricas lingüísticas.

Métricas de correspondencias de cadenas		
Levenstein	BlockDistance	Jaro
JaroWinkler	JaccardSimilarity	ChapmanLengthDesviation
QGramsDistance	CosineSimilarity	DiceSimilarity
EuclideanDistance	MongeElkan	NeedlemanWunch
ChapmanMeanLength		

En la tercera fase se calcula la similitud entre las etiquetas; el uso de métricas de similitud de cadenas es un valor sintáctico que se le agrega al operador como una medida de error, que se puede utilizar en un análisis posterior en la búsqueda de conflictos si existen errores de escritura en el diseño de los esquemas o existen diferencias significativas entre las cadenas las cuales generan conflictos lingüísticos que deben resolverse. En este cálculo se realiza la combinación de correspondencias. Además, se le agregan unos valores de predicción como los descritos en (Gal, 2006). El filtrado de las relaciones tiene como objetivo obtener aquellas relaciones que presentan una relación fuerte y aquellas que presentan una relación débil. Entiéndase por relación fuerte la que presenta una relación semántica de equivalencia y el valor de similitud sintáctica por encima de un umbral especificado. Entiéndase por relación débil aquella que expresa la existencia de posibles conflictos entre los esquemas de datos. Las relaciones que cumplen con las especificaciones se almacenan en un objeto serializado para luego guardar los resultados en un documento XML, y las que están en un rango cercano al umbral o presenten algún otro conflicto se almacenan en otro documento XML para ser analizadas por otro módulo.

Resultados y discusión

El operador JMatcher extrae los grafos de los esquemas conceptuales MERODE y encuentra correspondencias entre los nodos de los grafos, seleccionando un conjunto de métricas lingüísticas para combinar sus resultados y formando complejos conceptos semánticos a partir de las etiquetas de los nodos. La salida del operador consiste en tres grupos de correspondencias. El primer grupo lo conforman las correspondencias que son validadas semántica y sintácticamente. La tabla 2 muestra las lista de métricas utilizadas en el cálculo de la correspondencia sintáctica. El segundo grupo se forma con las correspondencias que no son validadas y el tercer grupo contiene aquellas correspondencias que están en cercanas al rango de validación pero como no existe un criterio de decisión

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

implementado para definir si pertenece al primero o al segundo se mantienen en otro grupo. Los umbrales de selección escogidos fueron 0.3 de mínimo y 0.8 de máximo debido a que distribuyen bien las relaciones y no se agrupan grandes conjuntos de relaciones en una misma clasificación.

Un aspecto importante a tratar es el caso de desigualdad, la cual aparece casi siempre cuando las palabras tienen una relación directa, o sea cuando se cumple un tipo de sinonimia. Pero cuando la relación de sinonimia está dada por la inclusión de una cadena en otra, el sentido de la desigualdad se pierde y la relación desaparece, es decir solo existe la relación de equivalencia.

Al comparar los esquemas (véase anexo 1 y 2), se obtuvieron 688 relaciones semánticas, de ellas 525 son de menos generalidad, 118 son de equivalencia, 43 de mayor generalidad y 2 de desigualdad. El agrupamiento de las relaciones como resultado del operador da 40 posibles relaciones de correspondencia al grupo fuerte, 110 de incertidumbre y de 487 son posibles errores como se muestra en la Figura 6.

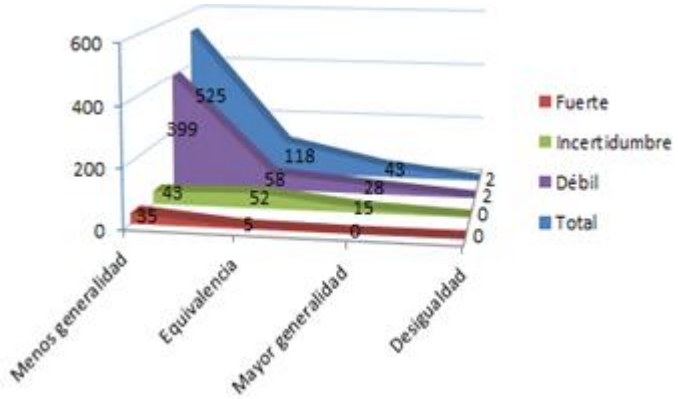

Figura 6. Distribución de las relaciones

La existencia de 35 relaciones de menos generalidad no expresa una correspondencia segura por lo que llevan un análisis de correspondencias estructurales para lograr una mayor especificación. La existencia de 58 relaciones de equivalencias como posibles errores indica que existen conceptos en los esquemas que se diseñaron empleando términos que son sinónimos en algún contexto.

Para la continuidad de este trabajo se propone implementar otras técnicas de agregación al combinar los resultados de similitud sintáctica basada en funciones de optimización y métodos estructurales que tengan en cuenta el tipo de construcciones utilizadas para representar los conceptos, con el objetivo de tener otro criterio para analizar las

relaciones de incertidumbre y poder identificar conflictos más complejos. También se propone mejorar el análisis de expansión de los acrónimos y el filtrado de las relaciones.

Conclusiones

El operador de correspondencia JMatcher encuentra las relaciones semánticas de equivalencia, desigualdad, mayor o menor generalidad. La combinación de los resultados se realiza en el cubo de similitud y se utilizan las estrategias para la construcción de la matriz de similitud. Los resultados de cada relación contienen valores de predicción para ayudar en la clasificación de las relaciones en fuertes o débiles. Aquellas que constituyen conflictos se almacenan como otra categoría porque llevan otro tipo de análisis.

Referencias

- BATINI C. y LENZERINI M., A comparative analysis of methodologies for database schema integration. ACM Computing Surveys. 1986. 18(4). p. 323-364.
- BERLIN J. y MOTRO A. Autoplex: Automated discovery of content for virtual databases. In Proc. Int. Conf. on Cooperative Information Systems. 2001. p.108–122.
- BERGAMASCHI S. y CASTANO S. Semantic integration of heterogeneous information sources. Data & Knowledge Engineering. 2001. 36(3) p.215–249.
- BERNSTEIN P. A. y MADHAVAN J. Generic Schema Matching, Ten Years Later. 2011.
- CASTANO S., De ANTONELLIS V. Global viewing of heterogeneous data sources. IEEETrans. Knowl and Data Eng. 2001. 13(2) p.277–297.
- CHENG R. y GONG J. Managing uncertainty of XML schema matching. In Proc.26th Int. Conf. on Data Engineering. 2010. p. 297–308.
- DOAN A., MADHAVAN J. Learning to map between ontologies on the semantic web. In Proc. 11th Int. World Wide Web Conf. 2002. p. 662–673.
- DONG X. L. y Halevy A. Y. Data integration with uncertainty. In Proc. 33rd Int. Conf. On Very Large Data Bases. 2007. p.687–698.
- GAL A., MODICA G. Automatic ontology matching using application semantics. AI Magazine 2005. 26(1).
- GAL A. y MARTINEZ M. V. Aggregate query answering under uncertain schema mappings. In Proc. 25th Int. Conf. on Data Engineering. 2009. p. 940–951.

- GAL, A. Managing uncertainty in schema matching with top-k schema mappings. Journal of Data Semantics 2006. 6.
- GIUNCHIGLIA F. y SHVAIKO P., S-Match: an Algorithm and an Implementation of Semantic Matching. 2004. p. 61-75.
- HONG H. D. y RAHM E. COMA - A system for Flexible Combination of Schema Matching Approaches.
- MADHAVAN J. y BERNSTEIN P. A. Generic Schema Matching using Cupid. 2001.
- MAGNANI M. y RIZOPOULOS N., Schema Integration Based on Uncertain Semantic Mappings. In Proc. 24th Int. Conf. on Conceptual Modeling. 2005. p. 31–46.
- MELNIK S., GARCÍA H. Similarity Flooding: A Versatile Graph Matching Algorithm. Extended Technical Report. 2002.
- MELNIK S. y RAHM E., A programming platform for generic model management. In Proc. ACM SIGMOD Int. Conf. on Management of Data: 2003. p. 193–204.
- MILLER R. J. y HAAS L. M. Schema mapping as query discovery. In Proc. 26th Int. Conf. on Very Large Data Bases. 2000. p.77-88.
- MILLER R .J. y HERNÀNDEZ M. A. The Clio project: Managing heterogeneity. SIGMOD Record 2001 30(1). p. 78–83.
- RAHM E. y BERNSTEIN P. A. On Matching Schemas Automatically. 2001.
- RAHM E., BERNSTEIN P. A. A survey of approaches to automatic schema matching. 2001.
- SALEEM K. y BELLAHSENE Z. Performance oriented schema matching In Proc.18th Int. Conf. Database and Expert Systems Appl. 2007. p. 844–853.
- SHETH A., LARSON J. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput. Surv. 1990. 22(3). p. 183–236.
- SNOECK M., et al. Object Oriented Enterprise 2011.
- SHVAIKO P. y EUZENAT J. Asurvey of schema-based matching approaches. Journal of Data Semantics. 2005. 4. p. 146 – 171.
- TROYANO, J. A. "WordNet."

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 32-47
http://rcci.uci.cu

Anexos

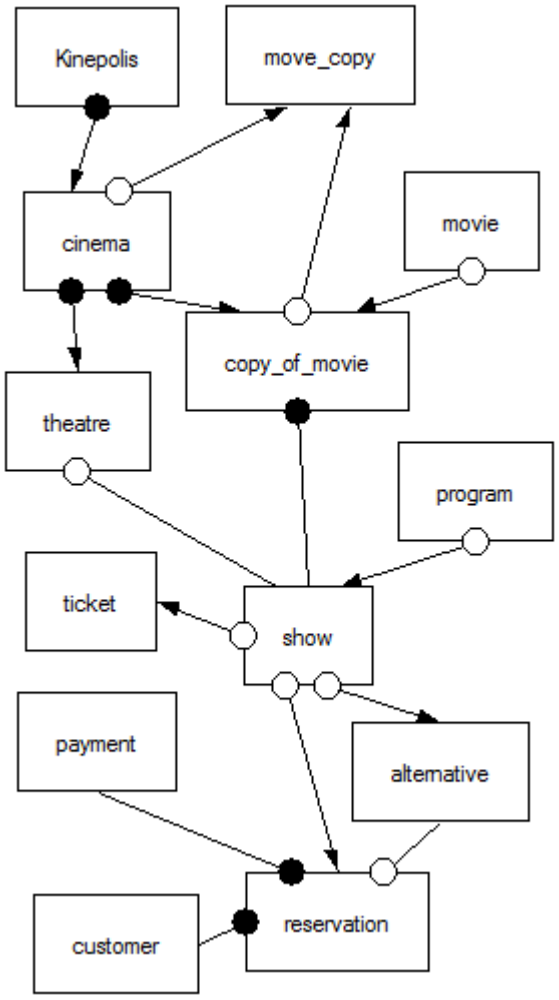

Figura 8. Kinepolis 8.

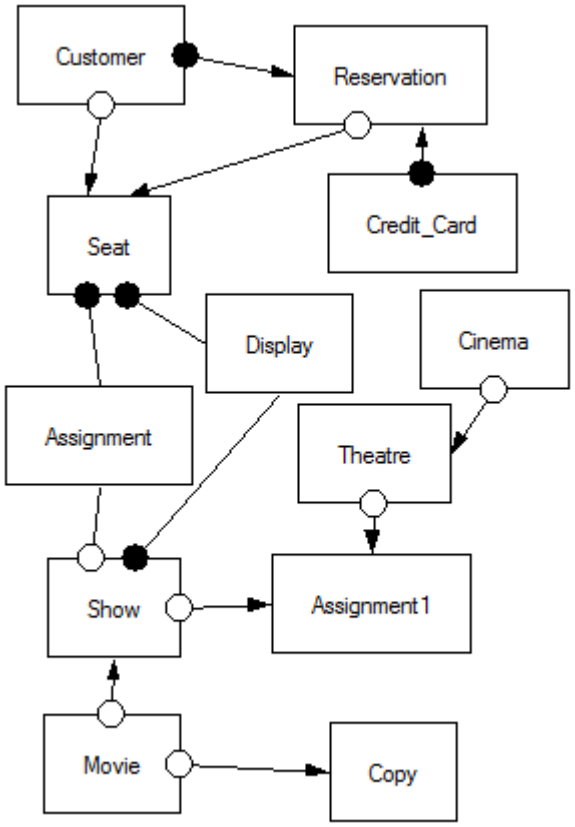

Figura 9. Kinepolis 10.