Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Cadena Martínez, Rodrigo; Quintero Téllez, Rolando; Moreno Ibarra, Marco Antonio; Torres Ruiz, Miguel; Guzmán Lugo, Giovanni

Comparación semántica de conjuntos de datos geográficos conceptualizados por medio de ontologías

Computación y Sistemas, vol. 17, núm. 4, octubre-diciembre, 2013, pp. 569-581
Instituto Politécnico Nacional
Distrito Federal, México

Comparación semántica de conjuntos de datos geográficos conceptualizados por medio de ontologías

Rodrigo Cadena Martínez, Rolando Quintero Téllez, Marco Antonio Moreno Ibarra,
Miguel Torres Ruiz y Giovanni Guzmán Lugo

Depto. Procesamiento Inteligente de Información Geoespacial,
Centro de Investigación en Computación, Instituto Politécnico Nacional,
México

{rcadenam, quintero, marcomoreno, mtorres, jguzmanl}@cic.ipn.mx

Resumen. En este artículo se presenta una metodología para la comparación de conjuntos de datos geográficos (CDG) con un enfoque semántico. Esto se realiza a través de su representación conceptual, es decir, mediante ontologías. Se propone de esta forma debido a que la mayoría de las comparaciones que se realizan es a través de técnicas sintácticas y no a través de la semántica de los conceptos. La metodología se compone de cuatro etapas: conceptualización, en la cual se genera una ontología de aplicación, por medio de los metadatos de los CDG a comparar. La etapa es el enriquecimiento, utilizada para poblar la ontología con datos, con la finalidad de generar una ontología enriquecida por cada CDG a comparar. En la etapa de comparación, se implementa un conjunto de algoritmos para medir la similitud entre las propiedades y relaciones de las ontologías enriquecidas. Finalmente, la etapa de interpretación y representación se muestran los resultados obtenidos de la comparación. Como caso de estudio se realizó la comparación de CDGs proporcionados por dos instituciones públicas de México.

Palabras Clave. Comparación, ontología, enriquecimiento, alineación, semántica.

Ontology-Driven Semantic Comparison between Geographic Data Sets

Abstract. In this paper a methodology for making the comparison and assessment of geographic datasets (CDG) with a semantic approach is presented. It is carried out by means of the processing of conceptual representations, in a particular case with ontologies. It is important to mention that the most of works related to CDG comparison use syntactic approaches for the analysis, according to above we propose to change this focus with semantic approaches. The methodology is composed of four stages. Conceptualization, where by using metadata from the CDGs the application ontology is generated. The enrichment stage consists in populating the ontology with data in order to create an enriched ontology for each CDG to compare. At the comparison stage, a set of algorithms for measuring the similarity between the properties and relationships of the enriched ontologies is developed. Finally, the interpretation and representation stages are in charge of showing the results of the CDGs comparison. As a case study, two Mexican government institutions provided the CDGs in order to make the comparison and assessment.

Keywords. Comparison, ontology, enrichment, ontology align, semantics.

1 Introducción

El uso cada vez más extendido de los Sistemas de Información Geográfica (SIG*),* ha conducido a una mayor utilización de los datos geoespaciales; lo cual propicia que dichos datos se compartan, se consulten y se intercambien cada vez con mayor frecuencia, llegando a ser utilizados para propósitos distintos para las cuales fueron generados [1]. Para asegurar que los datos puedan ser utilizados, tanto por sistemas informáticos como por los propios usuarios, es necesario que éstos sean "documentados correctamente" [2]. Por ello, es necesario describir detalladamente los elementos que los conforman para lograr que se mejore el

ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-009

acceso e intercambio a los mismos [3]. Hoy en día, existe una gran variedad de datos, provenientes de diversos productores, generados, utilizando diferentes tecnologías (Percepción Remota, fotogrametría, GPS, etc.), lo que origina una diversidad en sus características [2]. Lo anterior resulta evidente al integrar o fusionar datos provenientes de distintas fuentes, en donde se hacen evidentes las diferencias en la forma en la que éstos fueron conceptualizados [4]. Esta problemática da una idea del por qué resulta difícil seleccionar los datos adecuados para una aplicación SIG en particular [5]. A pesar de que existen algunos esfuerzos por estandarizar o unificar los criterios sobre cómo almacenar y representar los datos geográficos, esto no se ha logrado totalmente. Por ejemplo, la creación de una infraestructura de datos espaciales, a través de la implementación de un soporte informático tecnológico que actúa como una red de servicios SIG [6]. No obstante, no se ha logrado conseguir una estandarización, y cada productor de datos geográficos es libre de seguir cualquier estándar o realizar el suyo propio de acuerdo con sus necesidades específicas. Considerando lo anterior, se puede decir que un usuario interesado en identificar si los datos geográficos disponibles son útiles para sus fines, se encontrará ante algunas dificultades, las cuales involucran aspectos del diseño de los datos. En este sentido, una herramienta para comparar conjuntos de datos geográficos (CDG) resulta útil, ya que permite evaluar los aspectos relacionados con la utilidad de los datos. En este artículo se propone una metodología para realizar esta comparación, considerando diferentes CDGs para determinar el grado de similitud que tienen con base en su contenido semántico.

La metodología propuesta considera aspectos descriptivos (atributos), además de geográficos (relaciones espaciales), por lo que se propone utilizar una representación de los datos capaz de incluir lo necesario para compararlos de forma adecuada. Por ello, se incorporan las propiedades de los datos, mediante una estructura de representación del conocimiento que permite integrar todos los aspectos relevantes [7]. Para esto se propone el uso de ontologías, las cuales permiten representar los aspectos más significativos del dominio de los datos geográficos (clases de objetos, relaciones, restricciones y propiedades) [8]. Además de que constituye una representación formal del conocimiento que se tiene sobre el dominio de interés [9].

En el contexto de bases de datos, en [10] se muestra como se integran múltiples bases de datos individuales en una sola que contiene la información de todas las bases con las que fue diseñada y además existe un mapeo entre cada base de datos individual y la base de datos unificada. Por otra parte, con un enfoque también basado en bases de datos, pero con la diferencia de utilizar especificaciones textuales (metadatos, diccionarios de datos, entre otras), en [11] se presenta una infraestructura más completa para la construcción de bases de datos, considerando clases y conceptos para tal fin; en dicha propuesta se requiere de la asistencia de un experto. Además, se llega a la conclusión de que para extender este proceso es necesario completar la ontología con propiedades, relaciones y definiciones. Posteriormente, se realiza la comparación de las ontologías, para lo cual se utiliza una técnica llamada *TaxoMap* [12], en la que se clasifican las etiquetas de las ontologías que se compararán, en dos tipos: el primero constituido por verbos, adverbios y adjetivos; y el segundo conformado por artículos y pronombres [13]. La aportación de este trabajo radica en que se realiza una buena descripción de los datos para poder ser almacenados en una base de datos y posteriormente realizar su comparación, pero le hace falta incluir propiedades y relaciones que enriquecerían la comparación entre los datos.

Un trabajo orientado hacia la medición de interoperabilidad entre geo-ontologías se describe en [8], resaltando la importancia de la interoperabilidad entre diversas ontologías con diferentes bases de datos, para poder compartir información. Además, se menciona que esta medición se basa en el modelo algebraico para la descripción de metadatos creado por Bernstein [14], con la intención de llegar a términos más generales y que se presente una mayor probabilidad de que éstos se refieran a lo mismo.

El artículo está organizado como sigue: en la Sección 2 se describe la metodología de comparación de CDGs. En la Sección 3 se

presentan los resultados obtenidos sobre el caso de estudio y finalmente las conclusiones son descritas en la Sección 4.

2 Metodología propuesta

La metodología propuesta está compuesta de cuatro etapas: la *Conceptualización*, el *Enriquecimiento*, la *Comparación*, así como la *Interpretación y Representación de los resultados*. En la Figura 1 se muestra un diagrama a bloques de las etapas, así como las entradas y salidas de cada una de ellas.

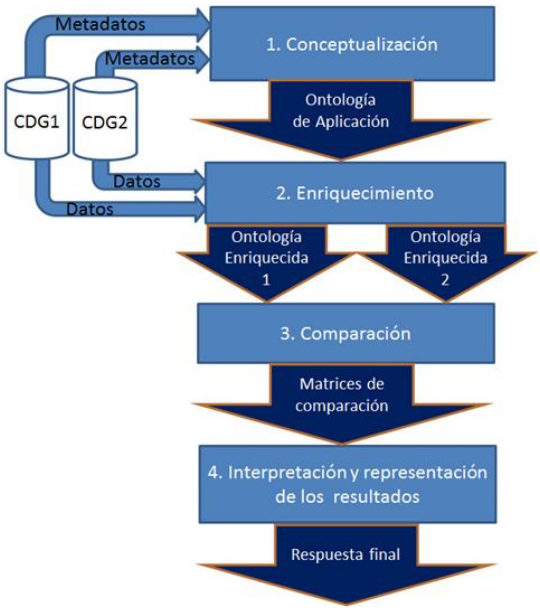

Fig. 1. Diagrama a bloques de la metodología general

2.1 Conceptualización

Esta etapa consiste en conceptualizar (describir) el dominio específico al que corresponden los CDG y plasmar esta conceptualización en una ontología de aplicación. Esta labor debe ser realizada por un experto, debido a que se requiere un análisis cognitivo para la interpretación de los metadatos y su posterior representación en la ontología de aplicación.

La ontología de aplicación describe las propiedades y las relaciones que se presentan en los CDG a comparar. El propósito de esta etapa es generar una representación formal del dominio al que pertenecen los datos geográficos que se van a comparar, por lo que se deben analizar los elementos de interés: conceptos, propiedades y relaciones. Esto se debe a que los elementos a identificar provienen de distintos CDG, y es necesario determinar si éstos pueden alinearse, es decir, establecer si los conceptos significan lo mismo.

2.1.1 Caracterización de propiedades

Considérense las siguientes definiciones:

Definición 1. $K(C, \Re, R)$ es la conceptualización del dominio, siendo C el conjunto de conceptos, $\Re$ el conjunto de los tipos de relaciones que existen entre los conceptos del dominio y R el conjunto de relaciones concretas entre los conceptos. Por otra parte, supongamos que A y B son los CDGs a comparar, entonces:

Definición 2. $S_P(x, y)$ es una función de similitud entre propiedades que establece si un par de propiedades (x e y) se consideran similares o no, la cual está definida como se muestra en la Ecuación 1.

$$S_P(x,y) = \begin{cases} verdadero\ si\ x\ es\ similar\ a\ y \\ falso\ de\ otra\ forma \end{cases} \quad (1)$$

Axioma 1. Propiedades alineables (P_{ab}). Sean $a \in A$ y $b \in B$ dos elementos de los conjuntos de datos, además, P_a el conjunto de propiedades del elemento a y P_b el conjunto de propiedades del concepto b, entonces P_{ab} es el conjunto de propiedades que se encuentran en p_a que tienen elementos similares en p_b. En otras palabras, se debe buscar qué propiedades representan la misma característica para sus respectivos conceptos. En caso de existir dicha similitud se considera tal propiedad como una *propiedad alineable* (ver Ecuación 2).

$$P_{ab} = \{p_a \in P_a, p_b \in P_b | S(p_a, p_b)\} \quad (2)$$

Axioma 2. Propiedades no alineables ($\bar{P}$). Dados los conjuntos P_a y P_b, las propiedades no

alineables son aquellas que no cumplen con el Axioma 1.

Axioma 3. Umbral de similitud de propiedades (U_{sim}). Es un valor que determina la cantidad mínima de propiedades alineables que deben de tener los conceptos para considerarse alineables.

Definición 3. Términos alineables (Ψ). Sean $a \in A$ y $b \in B$ dos elementos de los conjuntos de datos, entonces si el número de propiedades alineables superan el umbral de similitud de propiedades, definido en el Axioma 3, se dice que los conceptos *a* y *b* son términos alineables (ver Ecuación 3).

$$\boldsymbol{\Psi(A,B) = \{a \in A, b \in B, |P_{ab}| > U_{sim}\}} \quad (3)$$

Definición 4. Términos no alineables ($\overline{\Psi}$). Sean $a \in A$ y $b \in B$ dos elementos de los conjuntos de datos, entonces si el número de propiedades alineables no superan el umbral de similitud de propiedades, definido en el Axioma 3, se dice que los conceptos *a* y *b* son términos no alineables.

Definición 5. Conceptos de la ontología de aplicación (C). Dados los conjuntos de términos alineables y no alineables, se determina el conjunto de conceptos alineables: si a y b son términos alineables, entonces estos forman el concepto alineable $c \in C$, si existen más términos alineables con a o con b también formarán al concepto alineable $c \in C$. Los conceptos no alienables son aquellos formados por un solo término no alineable. La unión de estos conjuntos genera los conceptos de la ontología de aplicación. (Ecuación 4)

$$C = \{a, b \in \Psi(A,B)\} \bigcup \{a, b \in \overline{\Psi}(a,b)\} \quad (4)$$

En resumen, los pasos a seguir para determinar los conceptos alineables y no alineables, se propone realizar una analogía al procedimiento definido en [18,19], en donde se define la similitud entre los conceptos de ambas conceptualizaciones. Dichos pasos son:

1. Cada elemento $a \in A$, se comparará con cada concepto $b \in B$.
   1.1 Cada propiedad del elemento a, se comparará con cada propiedad del elemento b.
      1.1.1 Si la cardinalidad del conjunto de propiedades alineables de a y b, superan el umbral de similitud U_{sim}, entonces se determina que a y b son instancias alineables.
      1.1.2 En caso de que el umbral de similitud U_{sim}, no sea superado, se determina que a y b son instancias no alineables.
2. El conjunto de conceptos C estará formado por la unión de los conceptos alineables (formados por las instancias alineables entre sí) y los conceptos no alineables (uno por cada instancia no alineable con alguna otra).

2.1.2 Caracterización de relaciones

El siguiente paso de la conceptualización consiste en identificar cómo se relacionan los conceptos alineables y no alineables. Entonces, todos los conceptos deben estar relacionados, las relaciones aportan información valiosa de los conceptos con respecto a su entorno. Dependiendo del dominio en el que se encuentren, existirán relaciones que definan su comportamiento y acciones en dicho dominio. El objetivo de este análisis es caracterizar estas relaciones para integrarlas en la ontología de aplicación.

Por cada concepto de la ontología de aplicación es necesario realizar un análisis, con la finalidad de determinar las relaciones que tiene con cada uno de los conceptos y además determinar de qué tipo es dicha relación. Los conjuntos que caracterizan las relaciones (R y $\Re$) se integrarán en la ontología de aplicación para completar los conceptos, introduciendo las relaciones que existen entre ellos. Estas relaciones son representadas en la ontología análogamente como son representadas las propiedades debido a que las relaciones al igual que las propiedades son inherentes a los conceptos.

Al finalizar la etapa de Conceptualización, se ha construido una ontología de aplicación, en la cual se encuentran todos los elementos de los

CDGs divididos en dos clasificaciones: los conceptos alineables y no alineables, esto se realiza con la intención de poder instanciar en estos conceptos cualquier dato contenido en los CDGs y además saber cuáles de ellos comparten una misma conceptualización.

2.2 Enriquecimiento de las ontologías

En esta etapa se realiza la representación conceptual de los conjuntos de datos, es decir, a partir de la base de datos que contiene objetos geográficos, se genera una ontología en la que cada elemento geográfico contenido en la base de datos se encuentra representado como una instancia en la ontología enriquecida. Esto se hace por medio de un proceso de instanciación o "poblado" de la ontología de aplicación. El primer paso para enriquecer la ontología consiste en establecer un mapeo entre cada uno de los elementos en el CDG y la ontología de aplicación.

Definición 6. Mapeo (m). Sean A un conjunto de datos geográficos y O una ontología. Se asigna para cada elemento $a \in A$, un elemento $m(a)$ que pertenece a la ontología O, es decir, un mapeo de A en O, es representado por $m: A \rightarrow O$. Como se muestra en la Ecuación 5.

$$\forall a \in A, \exists c \in O \ni m(a) \rightarrow c \quad (5)$$

Para realizar el mapeo, se parte del hecho de que a pesar de que los datos geográficos puedan ser distintos, la estructura ontológica es la misma; ya que como se mostró en la etapa de Conceptualización, los conceptos que se representan fueron alineados. Esto permite que puedan ser mapeados en la ontología de aplicación, que es común para ambas conceptualizaciones. En la Figura 2 se ilustra el mapeo de los elementos del primer CDG con la ontología de aplicación, análogamente los elementos del segundo CDG se mapean con la misma ontología.

En esta etapa se transforman los datos de una representación sintáctica a una representación semántica. Por lo tanto, estos datos adquieren un *significado* por estar vinculados en una estructura semántica que tiene la representación del dominio.

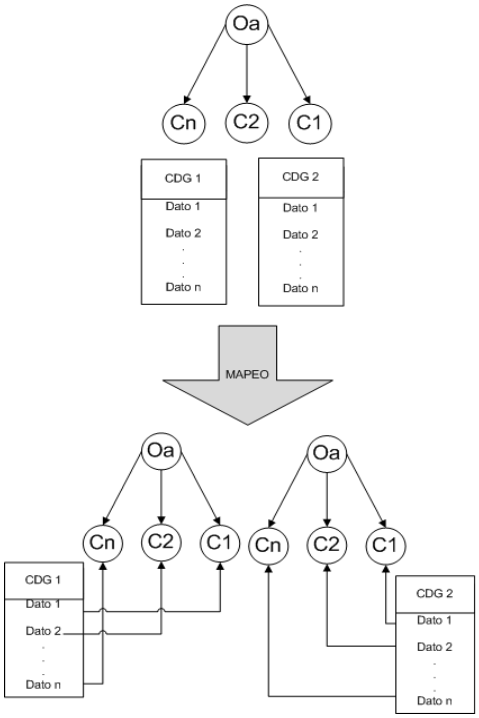

Fig. 2. Mapeo de los CDG y la ontología

Una vez establecido el mapeo, los pasos para enriquecer la ontología de aplicación son:

Para cada conjunto de datos geográfico CDG_x que se desean comparar, se crea la ontología enriquecida O_{ex} teniendo como base la ontología de aplicación creada en la etapa de conceptualización:

1.1 Para cada elemento $c \in CDG_x$, sea $m(c)$ un mapeo entre el objeto y la ontología de aplicación *O,* entonces:
- 1.1.1 Se instancia c en la ontología enriquecida Oe_x
- 1.1.2 Por cada propiedad p que pertenece al concepto c sea $m(p)$ un mapeo entre p y una propiedad del concepto de la ontología de aplicación *O*, entonces:
  - 1.1.2.1 Se instancia p en la ontología enriquecida Oe_x
- 1.1.3 Por cada relación r que pertenece al concepto c, sea $m(r)$ un mapeo entre r y una relación del concepto de la ontología de aplicación *O*, entonces:

1.1.3.1 Se instancia r en la ontología enriquecida Oe_x

Al finalizar esta etapa se ha construido una ontología enriquecida por cada uno de los CDG a comparar.

2.3 Comparación

En esta etapa se describe la comparación entre las instancias de las ontologías enriquecidas generadas en la etapa anterior. Por tanto, se propone comparar cada instancia de un concepto de una ontología enriquecida, contra las instancias del mismo concepto de la otra ontología enriquecida, para realizar la comparación se parte del hecho de que las ontologías enriquecidas tienen una estructura común debido a que fueron generadas con base en la misma ontología de aplicación que fue obtenida analizando los términos de los CDG y representan una conceptualización común de los CDGs.

Con base en lo anterior, es posible comparar las instancias de un concepto, con las instancias del mismo concepto en la otra ontología. Para realizar la comparación de las ontologías enriquecidas, se propone el siguiente procedimiento:

1. Se establecen la matriz de similitud de instancias (MSI) y la matriz binaria de correspondencia de instancias (MCI) como vacía.

2. Se recorre una de las ontologías enriquecidas (Oe_a) hasta encontrar un concepto (C_i) que tenga al menos una instancia (I_{1a}) y se comparará con todas las instancias $I_{Kb}, (K = 1, \dots, m)$ del mismo concepto (C_i) en la otra ontología enriquecida (Oe_b). Este paso se repetirá por cada instancia $I_{Ja}, (J = 1, \dots, n)$ de Oe_a.

3. El siguiente paso consiste en determinar el valor de la similitud entre cada par de instancias, mediante la comparación de cada una de las propiedades alineables de una instancia (I_{Ja}) con las propiedades alineables de las otras instancias (I_{Kb}).

   3.1 Se establece el valor de similitud a cero ($sim = 0$)

   3.2 Para cada propiedad alineable de la instancia I_{Ja} (p_{Jx}) se obtiene la propiedad alineable correspondiente en la instancia I_{Kb} (p_{Kx}).

   3.3 Si los valores de las propiedades son iguales se incrementa en 1 el valor de sim.

   3.4 Se establece el valor de similitud de las instancias como el valor de sim entre el número de propiedades *alineables*.

4. Como resultado del paso anterior se obtiene un vector de valores de similitud (VS_{JK}), el cual describe qué tan similar es la instancia I_{Ja} del concepto (C_i) de la ontología enriquecida (Oe_a), respecto a todas las instancias del mismo concepto en la otra ontología enriquecida (Oe_b). Entonces, se determina el mayor valor de similitud (el valor más grande en VS_{JK}) que indica cual instancia (I_{Kb}) en la ontología enriquecida Oe_b es el que debe corresponderse con la instancia (I_{Ja}) de la ontología enriquecida Oe_a. Puede darse el caso de que más de una instancia de Oe_b resulte con el valor máximo de similitud, por lo que se establece que la instancia I_{Ja} se corresponde con todas las instancias de Oe_b que generen el valor máximo. Se genera el vector binario de correspondencias (VC_{JK}) con valores de 1 en los máximos y de cero en los demás.

5. Se añade (VS) como una fila de la matriz global de similitud (MSI) y se añade (VC_{JK}) como una fila de la matriz binaria de correspondencia de instancias (MCI).

6. Repetir los pasos 2 a 5 hasta que se hayan analizado todas las instancias en Oe_a.

Algorithm 1 Compara Instancias(Oe_1,Oe_2)

Input: Dos ontologías enriquecidas: Oe_1 y Oe_2
Output: Matriz de Similitud de Instancias MSI y la matriz binaria de correspondencia de Instancias MCI

```
foreach Ii1 ∈ Cn ∈ Oe1 do
    max ← 0
    foreach Ij2 ∈ Cn ∈ Oe2 do
        sim ← 0
        foreach Px ∈ Ii1 do
            foreach Px ∈ Ij2 do
                if (Px ∈ Ii1) ≈ (Px ∈ Ij2) then
                    sim ← sim + 1
                end
            end
        end
        VSj ← sim / x
        if VSj > max then
            max ← VSj
        end
    end
    MSI ← VS
    foreach VSj do
        if VSj = max then
            VCj ← 1
        end
        else
            VCj ← 0
        end
    end
    MCI ← VC
end
```

Fig. 3. Algoritmo para comparar instancias y propiedades

En la Figura 3 se muestra el algoritmo propuesto.

Posteriormente, es necesario evaluar la similitud de las relaciones que existen en las dos ontologías enriquecidas para realizar este proceso es necesario analizar las Tablas de relaciones R_a y R_b, las cuales contienen todas las relaciones de las instancias de las ontologías enriquecidas. Los pasos se muestran a continuación:

1. Se establecen a cero los valores de la matriz de similitud de relaciones (MSR) de tamaño $|R_a| \times |R_b|$.
2. Por cada relación $\boldsymbol{r_a(I_{a1}, I_{a2}, r_a)}$ en la tabla de relaciones $(\boldsymbol{R_a})$ de la primera ontología enriquecida $(\boldsymbol{Oe_a})$ y por cada relación $\boldsymbol{r_b(I_{b1}, I_{b2}, r_b)}$ en la tabla de relaciones $(\boldsymbol{R_b})$ de la otra ontología enriquecida $(\boldsymbol{Oe_b})$, se establece que:

$$\boldsymbol{MSR(a,b) = \alpha * MSI(I_{a_1}, I_{b_1}) + \beta * MSI(I_{a_2}, I_{b_2}) + \gamma * S_R(r_a, r_b)}$$

donde: $\alpha + \beta + \gamma = 1$, S_R es la similitud entre las relaciones.

En la Figura 4 se muestra el algoritmo propuesto para la comparación de relaciones.

Algorithm 2 Compara Relaciones

Input: Las tablas de relaciones TR_1 , TR_2 y la matriz binaria de correspondencia de Instancias MCI
Output: Matriz de Similitud de Relaciones MSR

```
foreach Ra(Ia1, Ia2, ra) ∈ TR1 do
    foreach Rb(Ib1, Ib2, rb) ∈ TR2 do
        MSR(a,b) = α * MSI(Ia1, Ib1) + β * MSI(Ia2, Ib2) + γ * SR(ra, rb)
    end
end
```

Fig. 4. Algoritmo para comparar relaciones

Como resultado del algoritmo de la Figura 4, se ha generado una Matriz MSR, en la que se almacenan los valores de similitud entre todas las relaciones de una ontología enriquecida Oe_1, contra todas las relaciones de la otra ontología enriquecida Oe_2.

Al concluir esta etapa se han generado tres matrices. La primera MSI contiene todas las comparaciones de las instancias. La segunda MCI contiene las correspondencias de las instancias, es decir, almacena qué instancia de una ontología enriquecida es la más similar con la instancia de la otra ontología enriquecida. La última MSR almacena los valores de todas las comparaciones entre las relaciones.

2.4 Interpretación y representación de los resultados

Para interpretar y/o representar los resultados se proponen métricas estadísticas para llegar a una respuesta consistente; una para la comparación entre las instancias y otra para la comparación de las relaciones. Dichas métricas toman los rangos de sus valores debido a los algoritmos que se utilizaron para la obtención de las tablas MSI y MSR, en las cuales se define la similitud de las instancias y sus relaciones.

Para realizar las métricas de similitud se realiza una analogía con el trabajo realizado en [15] de tal forma que se proponen dos diferentes métricas, tanto para las instancias como para las relaciones como se muestra en la Figura 5, pero

como se menciona en [15] es posible cambiar tales valores por cualesquiera otros, de acuerdo con el dominio y a las necesidades específicas de cada aplicación.

En este artículo se propone la métrica de similitud de instancias definidas con los siguientes valores:

- Si, $MSI(a,b) = 0$ las instancias son completamente diferentes.
- Si $0 < MSI(a,b) \leq 0.35$, las instancias son poco similares.
- Si $0.35 < MSI(a,b) \leq 0.75$, las instancias son similares.
- Si $0.75 < MSI(a,b) \leq 1$, las instancias son muy similares.

La métrica de similitud de relaciones se propone con los siguientes valores:

- Si $MSI(a,b) = 0$, las relaciones son completamente diferentes.
- Si $0 < MSR(a,b) \leq 0.4$, las relaciones son poco similares.
- Si $0.4 < MSR(a,b) \leq 0.8$, las relaciones son similares.
- Si $0.8 < MSR(a,b) \leq 1$, las relaciones son muy similares.

Al finalizar esta etapa se proponen métricas para determinar la similitud entre las instancias y propiedades de los conceptos pertenecientes a las ontologías enriquecidas que se han comparado.

3 Resultados

En esta sección se presentan los resultados obtenidos con la implementación de la metodología propuesta para realizar la comparación de conjuntos de datos geográficos, utilizando una representación conceptual. El caso de estudio seleccionado fue la comparación de CDG de las vialidades que elaboraron dos instituciones gubernamentales de México: el Instituto Nacional de Estadística, Geografía e Informática, y el Instituto Federal Electoral.

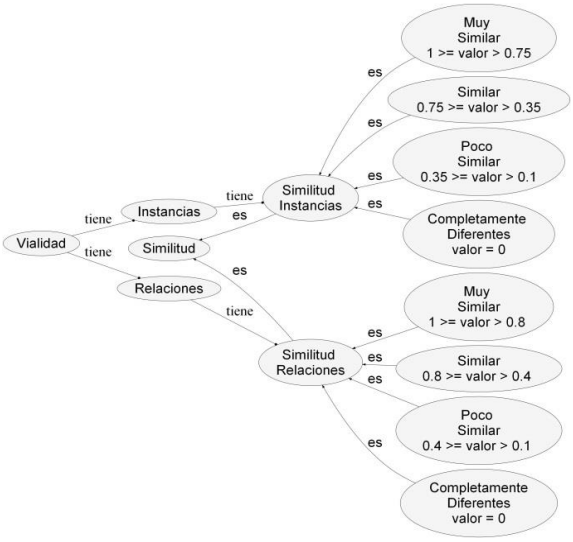

Fig. 5. Métricas de similitud de instancias y propiedades

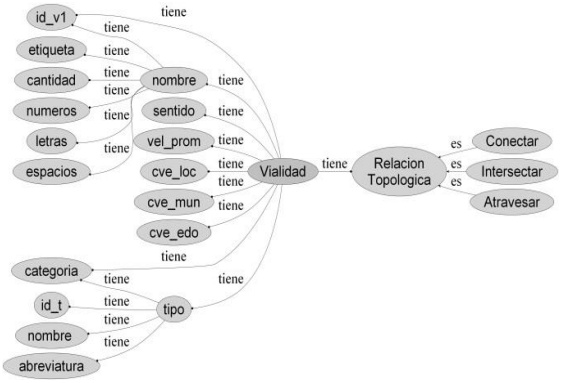

Fig. 6. Ontología de aplicación implementada

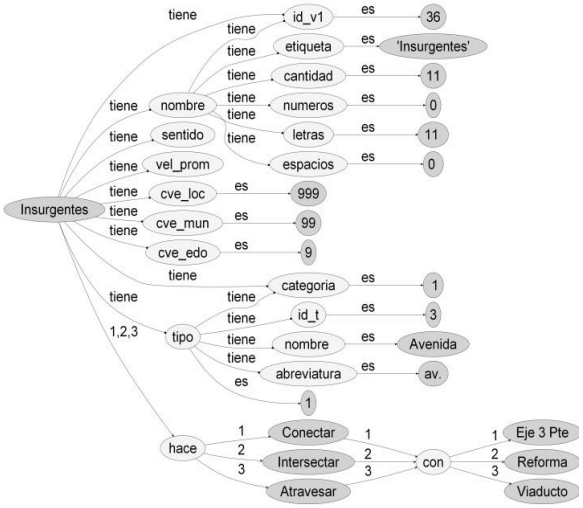

Fig. 7. Extracto de una ontología enriquecida

Computación y Sistemas Vol. 17 No. 4, 2013 pp. 569-581
ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-009

3.1 Resultados de la conceptualización

Como se mencionó anteriormente, el objetivo de la conceptualización es generar una ontología de aplicación que describa el dominio de los datos a comparar, la cual contiene todos los conceptos y propiedades de los CDG. Para el caso de estudio se realizó la ontología de aplicación para las vialidades, sus propiedades y relaciones, como se ilustra en la Figura 6.

La ontología de aplicación se genera por medio de la conceptualización de los metadatos que describen los CDGs a comparar, el objetivo de la ontología es entonces almacenar las vialidades de cualquiera de los CDGs mencionados anteriormente. Los datos con los que se poblará este concepto "*vialidad*", son los especificados en las bases de datos de los CDGs y deberán ser introducidos con sus propiedades y las relaciones que tengan con otras vialidades.

Las relaciones "*tiene*" se refieren a las propiedades que cada vialidad posee, tales como: su nombre, su sentido, su tipo, etc. Para la elaboración de la ontología nos basamos en la metodología GEONTO-MET [16], en la que se transforman relaciones por conceptos con la finalidad de enriquecer la ontología sin aumentar el número de relaciones axiomáticas en la conceptualización. Por ejemplo, para definir la relación "*conectar*" se utiliza la relación axiomática de acción "*hace*", junto con una relación de causalidad "*con*" para expresar la acción que realiza la vialidad; en este caso "*conectar*" y se emplea de la siguiente forma: **"Vialidad1 hace conectar con Vialidad2"**. Con este formato se introducirán las relaciones que tengan una vialidad relacionada con respecto a otras, en la siguiente sección se presenta un ejemplo que ilustra este procedimiento.

3.2 Resultados del enriquecimiento

El objetivo de la etapa de enriquecimiento es, como ya se ha mencionado, el de realizar la representación conceptual de los conjuntos de datos. Esto se realiza mapeando los conceptos instanciables de la ontología de aplicación con los datos contenidos en los conjuntos de datos (las instancias de dichos conceptos), esto significa que se genera una ontología enriquecida por cada conjunto de datos. Dicha ontología contiene instancias, y dichas instancias corresponden a la

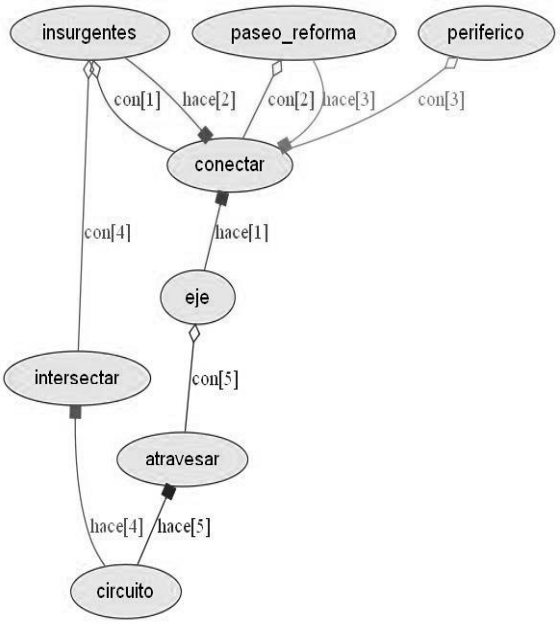

Fig. 8. Ontología enriquecida A, generada en la aplicación: *R-ontology*

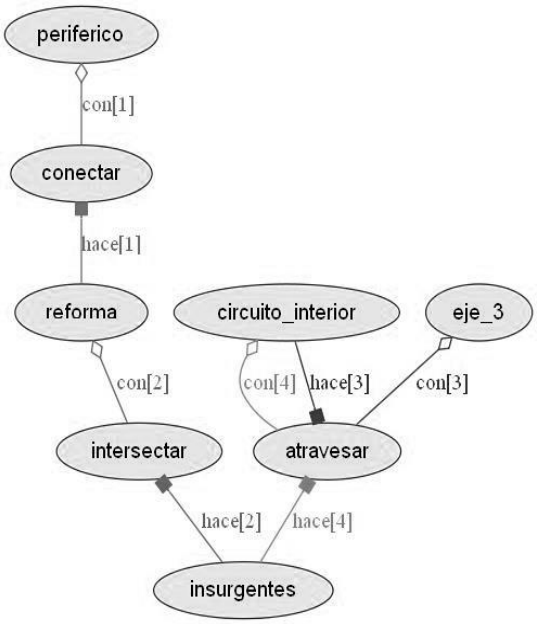

Fig. 9. Ontología enriquecida B, generada en la aplicación: *R-ontology*

información de los CDGs. Un extracto de una de estas ontologías enriquecidas se muestra en la Figura 7, en donde se resalta en color gris las instancias que son generadas por los datos de la vialidad específica “Insurgentes”.

Como parte de la etapa de enriquecimiento, se deben describir también las relaciones concretas que tienen las instancias (los datos) con las que se ha poblado la ontología de aplicación. Para este propósito se implementó una aplicación denominada *R-ontology* en donde se lleva a cabo este procedimiento de enriquecimiento.

En la ontología enriquecida (A), se introdujeron los siguientes datos experimentales. Los conceptos son: {*insurgentes, paseo_reforma, eje, circuito, periférico*}, las relaciones que tienen los datos describen como se encuentran interrelacionados objetos geográficos y estas son: {*intersecta, atraviesa, conecta*}. De esta manera se describen las relaciones concretas entre los conceptos, por ejemplo: {[*insurgentes hace intersectar con paseo_reforma*], [*paseo_reforma hace conectar con periférico*], [*circuito hace intersectar con insurgentes*], [*circuito hace atravesar con eje*], [*eje hace conectar con insurgentes*]}. La ontología enriquecida con sus relaciones se muestra en la Figura 8.

En la segunda ontología enriquecida (B), se introdujeron los siguientes datos experimentales: {*insurgentes, reforma, eje_3, circuito_interior, periférico*}, y las relaciones entre los conceptos son: {[*insurgentes hace intersectar con reforma*], [*insurgentes hace atravesar con circuito_interior*], [*reforma hace conectar con periférico*], [*circuito_interior hace atravesar con eje_3*]}. La ontología enriquecida B con sus relaciones se presenta en la Figura 9.

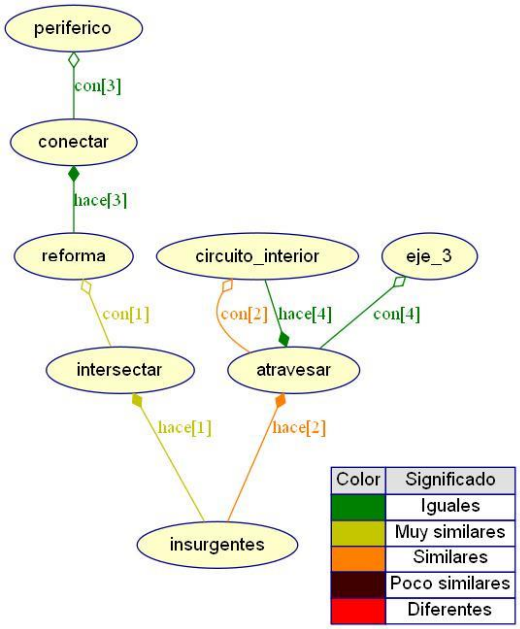

Fig. 10. Interpretación de la comparación de los CDGs a través de sus relaciones y propiedades

3.3 Resultados de la comparación

En la aplicación *R-ontology* se implementaron los algoritmos de comparación de instancias y de relaciones. En la Tabla 1, se muestran la Matriz de Similitud de Instancias (MSI), en donde se ilustran los resultados de la comparación entre todas las instancias de la ontología enriquecida (A) contra todas las instancias de la ontología enriquecida (B).

Tabla 1. Comparación entre todas las instancias

MSI(A,B)	Insurgentes	Reforma	Eje_3	Circuito interior	Periférico
Insurgentes	0.7500	0.6363	0.2500	0.2500	0.5000
Paseo_ reforma	0.6343	0.725	0.2500	0.2500	0.5288
Eje	0.2500	0.2500	0.9750	0.3750	0.3750
Circuito	0.2500	0.2500	0.3750	0.9750	0.6375
Periférico	0.5000	0.5125	0.2500	0.3823	0.7500

Tabla 2. Instancias más similares entre las dos ontologías enriquecidas

MCI(A,B)	Insurgentes	Reforma	Eje_3	Circuito_ interior	Periférico
Insurgentes	**1**	**0**	**0**	**0**	**0**
Paseo_ reforma	**0**	**1**	**0**	**0**	**0**
Eje	**0**	**0**	**1**	**0**	**0**
Circuito	**0**	**0**	**0**	**1**	**0**
Periférico	**0**	**0**	**0**	**0**	**1**

Tabla 3. Comparación de las relaciones existentes en ambas ontologías

MSR(A,B)	1	2	3	4
1	0.8 (muy similares)	0.5 (similares)	0.4 (similares)	0.2 (poco similares)
2	0.2 (poco similares)	0.2 (poco similares)	1.0 (muy similares)	0.2 (poco similares)
3	0.7 (similares)	0.2 (poco similares)	0.2 (poco similares)	0.5 (similares)
4	0.2 (poco similares)	0.7 (similares)	0.2 (poco similares)	1.0 (muy similares)
5	0.2 (poco similares)	0.2 (poco similares)	0.4 (similares)	0.2 (poco similares)

En la Tabla 2 se muestra la Matriz binaria de Correspondencia de Instancias (MCI), en donde se define con el valor de 1 a las instancias que sean más similares entre las ontologías.

En la ontología enriquecida (A) existen 5 relaciones y en la ontología enriquecida (B) se tienen 4. Por tanto, la cardinalidad de la tabla MSR es de (5x4). En la Tabla 3, se muestra dicha comparación.

3.4 Resultados de la interpretación y representación

La última fase consiste en vincular las dos matrices tanto de similitud de instancias como la de similitud relaciones, obtenidas en la fase de comparación. Para realizar este enlace es necesario comparar las relaciones entre las instancias del CDG (A) y el CDG (B), pero tomando las instancias más similares entre ambos CDGs como parámetro de entrada para la comparación. Con el objetivo de establecer una métrica global que determine que tan similares son los dos CDGs. Dicha métrica está conformada por los siguientes elementos: *iguales, muy similares, similares, poco similares y diferentes*, los cuales determinarán la respuesta final de que tan parecidos son los CDGs, a través de las propiedades y relaciones de las instancias. Las métricas son obtenidas en el sistema *R-ontology.* Por tanto, En la Figura 10, se muestra un ejemplo relacionado con la interpretación de la similitud de una manera gráfica.

4 Conclusiones

En este artículo se presenta una metodología para realizar la comparación basada en ontologías, y particularmente de conjuntos de datos geográficos. A diferencia de otros enfoques, el propuesto permite agrupar las diferentes propiedades y relaciones de los datos, siempre y cuando éstos sean conceptualizados e incluidos en una ontología de aplicación. Lo anterior resuelve el problema de la comparación únicamente desde el punto de vista sintáctico y léxico, debido a que se incluye la componente semántica en ella.

Debido al estudio y análisis de los metadatos, se llegó a la conclusión de que es posible conceptualizar dos conjuntos de datos

ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-009

geográficos siempre y cuando pertenezcan al mismo dominio. Dicho procedimiento se puede establecer siguiendo una serie de pasos definidos en este artículo.

El mapeo de conceptos, propiedades y relaciones de los datos es posible, si se define una ontología de aplicación previamente, la cual describe la semántica existente en el dominio de los conjuntos de datos.

El objetivo de realizar una comparación con base en las propiedades y relaciones consiste en obtener la similitud semántica de los conjuntos de datos a comparar. Para comprobar la utilidad de la metodología, se ha desarrollado un sistema (*R-ontology*) aplicado a un caso de estudio particular. Con ello, hemos obtenido resultados aceptables que nos indican que la metodología es viable y aplicable para casos que cumplan con los requisitos previos y establecidos en cuanto a características de los datos y entradas a los procedimientos definidos en la metodología.

Agradecimientos

Este trabajo ha sido desarrollado con el apoyo del Instituto Politécnico Nacional (IPN), por medio de los proyectos auspiciados por la Secretaría de Investigación y Posgrado (SIP): 20131215, 20130345, 20130347, así como del Consejo Nacional de Ciencia y Tecnología (CONACyT) mediante el proyecto 106692.

Referencias

1. **Sheeren, D., Mustière, S., & Zucker, J.D. (2009).** A data-mining approach for assessing consistency between multiple representations in spatial databases. *International Journal of Geographical Information Science,* 23(8), 961–992.
2. **Fisher, P., Comber, A., & Wadsworth, R. (2009).** What's in a name? Semantics, standards and data quality. *Spatial Data Quality*, 3–16.
3. **Fonseca, F.T., Egenhofer, M.J., Agouris, P., & Câmara, G. (2002).** Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS,* 6(3), 231–257.
4. **Fonseca, F. (2008).** Ontology-Based Geospatial Data Integration. *Encyclopedia of GIS* (812–815). New York: *Springer-Verlag.*
5. **Uitermark, H.T., Van Oosterom, P.J.M., Mars, N.J.I., & Molenaar, M. (2005).** Ontology-based integration of topographic data sets. *International Journal of Applied Earth Observation and Geoinformation,* 7(2), 97–106.
6. **Gould, M., Bernabé, M.A., Baes, J.A., Muro-Medrano, P.R., & Zarazaga, F.J. (2001).** Aspectos tecnológicos de la creación de una Infraestructura Nacional Española de Información Geográfica. *Mapping*, 67, 68–77.
7. **Raskin, R. (2008).** Knowledge Representation, Spatial. *Encyclopedia of GIS (*603–604). New York: *Springer-Verlag.*
8. **Fonseca, F., Camara, G., & Monteiro, A.M. (2006).** A Framework for measuring the interoperability of Geo-Ontologies. *Spatial Cognition and Computation*, 6(4), 309–331.
9. **Guarino, N. (1995).** Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies,* 43(5-6), 625–640.
10. **Partridge, C. (2002).** *The role of ontology in integrating semantically heterogeneous databases* (05/02). Padova, Italy: *LADSEB-CNR.*
11. **Mustière, S., Reynaud, C., Safar, B., & Abadie, N. (2009).** *Same words? Same worlds? Comparing ontologies underlying geographic data* (1521). LRI - Université Paris-Sud.
12. **Hamdi, F., Zargayouna, H., Safar, B., & Reynaud, C. (2008).** TaxoMap in the OAEI 2008 Alignment Contest. *Proceedings of Ontology Matching Workshop of the 7th International Semantic Web Conference.*
13. **Schmid, H. (1994).** Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing,* Manchester, UK.
14. **Bernstein, P. (2003).** Applying Model Management to Classical Meta Data Problems. *1st Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA,* 209–223.
15. **Quintero, R., Torres, M., Menchaca-Mendez, R., Moreno-Ibarra, M., Guzman, G., & Cortez, P. (2012).** Specialization of the Geomorphometric Concepts in Digital Elevation Models. *International Journal of Combinatorial Optimization Problems and Informatics,* 3(2), 44–53.
16. **Torres, M., Quintero, R., Moreno-Ibarra, M., Menchaca-Mendez, R., & Guzman, G. (2011).** GEONTO-MET: An approach to conceptualizing the geographic domain. *International Journal of Geographical Information Science,* 25(10), 1633–1657.

Rodrigo Cadena Martínez Doctor en Ciencias de la Computación por el Centro de Investigación en Computación. Actualmente es Director de carrera de Tecnologías de la Información de la Universidad Tecnológica Nacional Campus Marina-Cuitláhuac. Sus áreas de interés comprenden el procesamiento semántico y la comparación de conjuntos de datos geográficos.

Rolando Quintero Tellez Doctor en Ciencias de la Computación por el Instituto Politécnico Nacional de México. Es investigador en el Centro de Investigación en Computación del Instituto Politécnico Nacional. Autor de alrededor de 80 artículos científicos en congresos y revistas especializadas. Ha dirigido proyectos de investigación básica y de desarrollo tecnológico con instituciones nacionales. Áreas de interés representación del conocimiento, procesamiento de información semántica.

Marco Antonio Moreno Ibarra Investigador del Centro de Investigación en Computación del Instituto Politécnico Nacional. Autor de más de 80 publicaciones en congresos y revistas especializadas. Ha dirigido proyectos de investigación básica y de desarrollo tecnológico con instituciones como el Instituto de Ciencia y Tecnología del D.F. Fue Director de Sistemas, Informática y Telecomunicaciones del CONACyT en 2012. Áreas de interés, diseño de GIS, Similitud geoespacial, Generalización Cartográfica y GIS Voluntarios.

Miguel Torres Ruiz Profesor–Investigador del Centro de Investigación en Computación, del IPN. Asimismo, es Jefe del Laboratorio de Procesamiento Inteligente de Información Geoespacial. Ha publicado alrededor de 100 artículos científicos y de divulgación en diversos congresos, así como en revistas especializadas, tanto nacionales como internacionales. Sus áreas de interés se enfocan en el desarrollo de técnicas para la conceptualización del dominio geográfico, diseño inteligente de GIS, integración y recuperación semántica de fuentes de datos heterogéneas.

Giovanni Guzmán Lugo Profesor–Investigador del Centro de Investigación en Computación, del IPN. Ha publicado alrededor de 50 artículos científicos y de divulgación en diversos congresos y revistas. Sus áreas de interés se enfocan en procesamiento de datos geoespaciales raster, análisis de imágenes y dispositivos móviles.

Article received on 06/06/2012, accepted on 24/09/2012.