Fusionando conocimiento con ontologías

Resumen

El ser humano agrega nuevos conocimientos a los que ya posee, tomando en cuenta información novedosa, detalles adicionales, mayor precisión, sinónimos, homónimos, redundancias, contradicciones aparentes e inconsistencias entre lo que ya sabe y el nuevo conocimiento que le llega. De esta manera, adquiere incrementalmente información manteniendo siempre una ontología consistente. En cambio, los algoritmos para fusionar dos ontologías carecían de tales características, siendo meramente editores de ontologías que requerían de una persona para resolver los detalles.

Se presenta un método (OM, Ontology Merging) con su algoritmo e implementación, para fusionar o unir ontologías (provenientes de documentos en la Web) en forma automática (sin intervención humana), considerando las inconsistencias, contradicciones y redundancias entre ambas ontologías, de forma que el resultado sea lo más cercano a la realidad. OM obtiene buenos resultados, al compararlos contra uniones efectuadas manualmente. El uso repetido de OM permite adquirir gran información sobre un mismo tema.

Abstract

A human being adds new knowledge to his mind, taking into account new information, additional details, better precision, synonyms, homonyms, redundancies, apparent contradictions and inconsistencies between what he knows and the new arriving knowledge. In this way, he incrementally acquires information keeping at all times a consistent ontology. In contradistinction, algorithms to fuse two ontologies lacked these features, merely being computer-aided editors that required a person to solve the details.

A method (OM, Ontology Merging), its algorithm and implementation are presented, to fuse or join two ontologies (coming from Web documents) in an automatic fashion (no human intervention), producing a third ontology, taking into account the inconsistencies, contradictions and redundancies among both ontologies, thus delivering a result close to reality. OM gets good results, when they are compared against fusions manually carried out. Repeated use of OM allows acquisition of much information about a given topic.

Palabras clave

Ontología, Inteligencia Artificial, Representación del conocimiento, Web semántica, Unión de ontologías.

1 Introducción

Una persona acumula información a lo largo de su vida al ir añadiendo conocimiento (conceptos, relaciones, valores típicos…) nuevo al que ya posee en su mente (en su "ontología" o estructura del conocimiento), identificando redundancias, información nueva, pequeñas contradicciones, contradicciones severas, sinónimos y antónimos, entre otros casos. Hasta ahora, la computadora podía hacer el mismo proceso (unir conocimientos provenientes de dos fuentes u ontologías distintas) usando un editor que le facilita la tarea a una persona, quien finalmente decide. *El problema a resolver es cómo automatizar esa unión.*

Se presenta un algoritmo (OM, Ontology Merging) y su implementación, para fusionar dos ontologías en forma automática, obteniendo una tercera, considerando las inconsistencias, sinonimias, grado de precisión, contradicciones y redundancias entre ambas, de tal manera que el resultado sea lo más cercano a la realidad. La ontología resultante puede convertirse en una ontología del conocimiento actual si se fusionan varias ontologías de propósito general y especifica.

Los **trabajos actuales** se muestran en la sección 1.3; nuestros **resultados** están en la sección 4.

Como **trabajo futuro** cercano, esta ontología unificada servirá para contestar preguntas no triviales. Se muestran ejemplos (secciones 3 y 4) del funcionamiento de OM, y los planes a futuro acerca de cómo enriquecerlo, tales como la creación de un sistema (al que llamaremos OM*) que tome documentos de texto, los convierta [Nery, P., 2007] a ontologías para que OM las funda o una, y así crear otra ontología más extensa conteniendo el conocimiento que existía en las ontologías fuente. El objetivo es poder contestar preguntas no triviales como las que se estudian en: [Botello, A., 2007] usando la ontología resultante. Este mismo contestador o razonador nos servirá para comprobar la información unida.

De esta manera los usuarios de OM* podrán obtener respuestas concretas de OM* y no solo pedirle (como se le pide a Wikipedia o a Google) que proporcione documentos sobre ciertos temas y el usuario haga la deducción.

La propuesta se parece al Proyecto CYC [Reed, S. L., and Lenat, D., 2002], que buscó, durante una década, fabricar (manualmente) una ontología del conocimiento común.

1.1 Propósito de la Inteligencia Artificial (IA)

El área de las ciencias computacionales encargada de la creación de software y hardware que tenga comportamiento inteligente es la Inteligencia Artificial. Su fin es lograr que las computadoras manejen la información y respondan de manera similar a la inteligencia a nivel humano.

En Internet, actualmente existe una gran cantidad de información en billones de documentos, entre los cuales hay páginas Web, documentos de texto, portales de servicios, música, fotografías, mapas, etc. La forma de accederlos es a través de buscadores (Google, Citeseer, por citar algunos), los que recuperan solo una mínima parte de la información disponible de la gran masa de conocimiento de Internet, porque la acceden de manera sintáctica (a través de etiquetas, palabras y frases); es decir a través de comparaciones lexicográficas. Además, la respuesta es una larga lista de documentos, no siempre apropiada, y la información que se busca debe todavía procesarse o deducirse por la persona, leyendo cada uno.

1.2 ¿Cómo hacerlo?

Una manera de obtener más rápidamente la información buscada es **extraer** de documentos pertinentes en la Web su conocimiento (por ejemplo, pasándolo a estructuras de datos llamadas ontologías), **fusionar** estos conocimientos en una ontología cada vez mayor, y **explotar** esta ontología grande mediante un razonador que conteste preguntas complejas.
El trabajo que se expone en este artículo se centra en la fusión del conocimiento de manera automática, y su organización en forma analizable para la máquina. Esta fusión debe considerar no solo la sintaxis de las palabras y frases sino la semántica (la vecindad entre las palabras solicitadas, las palabras parecidas, las sinónimas, las homónimas, etc.), por lo tanto, es necesario hacer que la **representación** a usarse considere estos elementos. Para la computadora, esta representación describe al mundo real. Parte importante de este trabajo es, pues, representar elementos textuales (documentos, páginas HTML) a través de estructuras de datos dinámicas y ricas, como son las ontologías.

1.3 Qué se ha hecho

A diferencia de la creación manual de ontologías, como en: [Reed, S. L., and Lenat, D., 2002], el equipo de trabajo de este artículo está atacando el problema por otro camino: la creación de la ontología del conocimiento de manera automática, obteniendo pedacitos de conocimiento (ontologías chicas) y uniéndolos con cuidado (filtrando las inconsistencias, uniendo sinónimos...).
El proyecto que aquí se presenta, OM (Ontology Merging o Fusión de Ontologías), no se limita a formar una ontología del algún tipo especial de conocimiento. Cualquier rama del conocimiento es susceptible de ser atacada con el procedimiento de fusión de OM.

La enciclopedia electrónica Wikipedia amasa el conocimiento mediante documentos escritos por humanos. OM trata de amasar el conocimiento mediante ontologías que OM fabrica. La información en Wikipedia se guarda como texto en un lenguaje natural, los usuarios colocan las relaciones (ligas) a otros temas (documentos). La cantidad de enlaces limita o enriquece el documento; la información no es capaz de enlazarse sola: requiere de un humano. Si hay información inconsistente y redundante, ésta se controla a través de quienes publican la información. Como se ve, en Wikipedia la introducción de nueva información (textos), nuevas ligas entre información existente y detección y resolución de contradicciones se hacen manualmente.

Los métodos actuales de fusión de ontologías son ayudados por un usuario. PROMPT expuesto en [Fridman, N., and Musen, M., 2000], Chimaera [McGuinness, D., Fikes, R., Rice, J., and Wilder, S., 2000], OntoMerge expuesto en [Dou, D., McDermott, D., and Qi, P., 2002], IF-Map e ISI expuesto en [referencia de Internet 1] requieren que el usuario resuelva los problemas presentados durante la fusión, otros como FCA-Merge [Stumme, G., Maedche, A., 2002] usan el Análisis Formal de Conceptos para la representación de sus ontologías, forzando a éstas ser consistentes. La mayoría de las ontologías en la Web suelen presentar inconsistencias. El algoritmo que está tomando ventaja en el proceso de fu-

sión es HCONE-merge [Kotis, K., and Vouros, G., Stergiou, K., 2006] pues utiliza la base de datos semántica WordNet [Fellbaum, C., 1999] como información intermediaria para la fusión, requiriendo menos del apoyo del usuario, lo cual significa un importante avance hacia la unión automática de ontologías.

La dinámica de OM es que la misma máquina vaya juntando ontologías pequeñas que encuentre, para ir formando (lo más coherentes entre sí) otras más grandes. ¡Interesante y no menos desafiante!

1.4 El propósito de fusionar conocimiento

OM va encaminado a la fusión de conocimiento a través de ontologías. Si cada elemento del conocimiento en Internet se traduce a ontologías, se tiene una información que sería más útil al usuario. Por ejemplo, una ontología sobre la vida de Albert Einstein, obtenida de digamos 50 biografías. Si el usuario solicita información sobre este personaje, bastaría con formular una pregunta a esta ontología grande.

Por ahora, la fusión que OM lleva a cabo se comprueba manualmente. En un futuro cercano, esta fusión se comprobará mediante un contestador de preguntas (en proceso de construcción [Botello, A., 2007]), da una respuesta congruente con la información fuente. Los documentos de Wikipedia y los que Google devuelve no pueden contestar preguntas, necesitan una persona que los lea y las conteste.

2. Elementos de OM

Se presenta el funcionamiento de OM a través de ejemplos (documentos y páginas Web) tomados de Internet. Las ontologías que se fusionan en los ejemplos se crearon manualmente (hay un trabajo en progreso [Nery, P., 2007] que se ocupa de la traducción de documentos a ontologías), y se fusionaron sin intervención humana mediante OM.

2.1 Definición de ontología

Las ontologías tienen una larga historia en filosofía (refiriéndose a la existencia). Por ello es llamada la teoría del ser, es decir, el estudio de todo lo que es: qué es, cómo es y cómo es posible. La ontología se ocupa de la definición del ser y de establecer las categorías fundamentales o modos generales de ser de las cosas a partir del estudio de sus propiedades [referencia de Internet 2].

2.1.1 Ontología desde el punto de vista de la Computación

La ontología tiene un sentido diferente a la que se usa en la filosofía pues es una herramienta para poder compartir y reutilizar conocimientos entre sistemas de Inteligencia Artificial, previamente es preciso definir un vocabulario común en el cual, se encuentre representado el conocimiento (dominio del discurso); en este caso, específicamente una ontología es el conjunto de definiciones, clases, relaciones, funciones y otros objetos de este dominio del discurso.

2.1.2 Definición de Ontología según Gruber

El término ontología es una representación o ***especificación formal*** (una estructura sintáctica) ***de una conceptualización*** (conjunto de conceptos) ***compartida*** (común a varias personas u otros programas) [Gruber, T., 1993].

2.1.3 Ontología desde el punto de vista formal:

Actualmente no existe una teoría satisfactoria que caracterice formalmente la definición de las ontologías, no obstante ello han habido algunas aproximaciones como en [Kalfoglou, Y., and Schorlemmer, M., 2002].

Para comprender la lógica de las ontologías descritas en este artículo, una ontología es una tupla O = (Ç, R) donde:

Ç es un conjunto de nodos (que denotan conceptos) de los cuales algunos de ellos son relaciones.

R es un conjunto de restricciones, de la forma (r; c_1; c_2;…; c_k) entre la relación r y los conceptos c_1 hasta c_k (se usa c minúscula para referirse a cada concepto del conjunto Ç y punto y coma para separar los argumentos). Ejemplos: (cortar; tijera; papel), (transcribir; impresora; documento; tinta). En ese ejemplo, los conceptos que también son relaciones son *cortar*, *transcribir*. Las relaciones no están limitadas a dos argumentos. Nótese que una ontología es un hipergrafo con nodos del conjunto Ç e hiper-relaciones R.

Hasta ahora, las ontologías se han propuesto para resolver los problemas (esto es, inducir a la máquina a un mejor *entendimiento* del lenguaje natural del ser humano) en la Web Semántica; para ello, se usan lenguajes de definición de ontologías para especificar la información de las páginas Web. El propósito es que estos lenguajes expresen la semántica que se requiere. En realidad, se verá que los lenguajes actuales cumplen parcialmente con este propósito. Otra de las contribuciones de este trabajo es proporcionar una representación para las ontologías que considera características nuevas en su representación.

La Figura 1 muestra una ontología representada como una estructura de datos graficada como una red semántica. Los elementos (llamados nodos o conceptos) se ubican de acuerdo a la relación entre ellos, a veces de lo general a lo particular, otras veces enlazados a través de sus características y tipos.

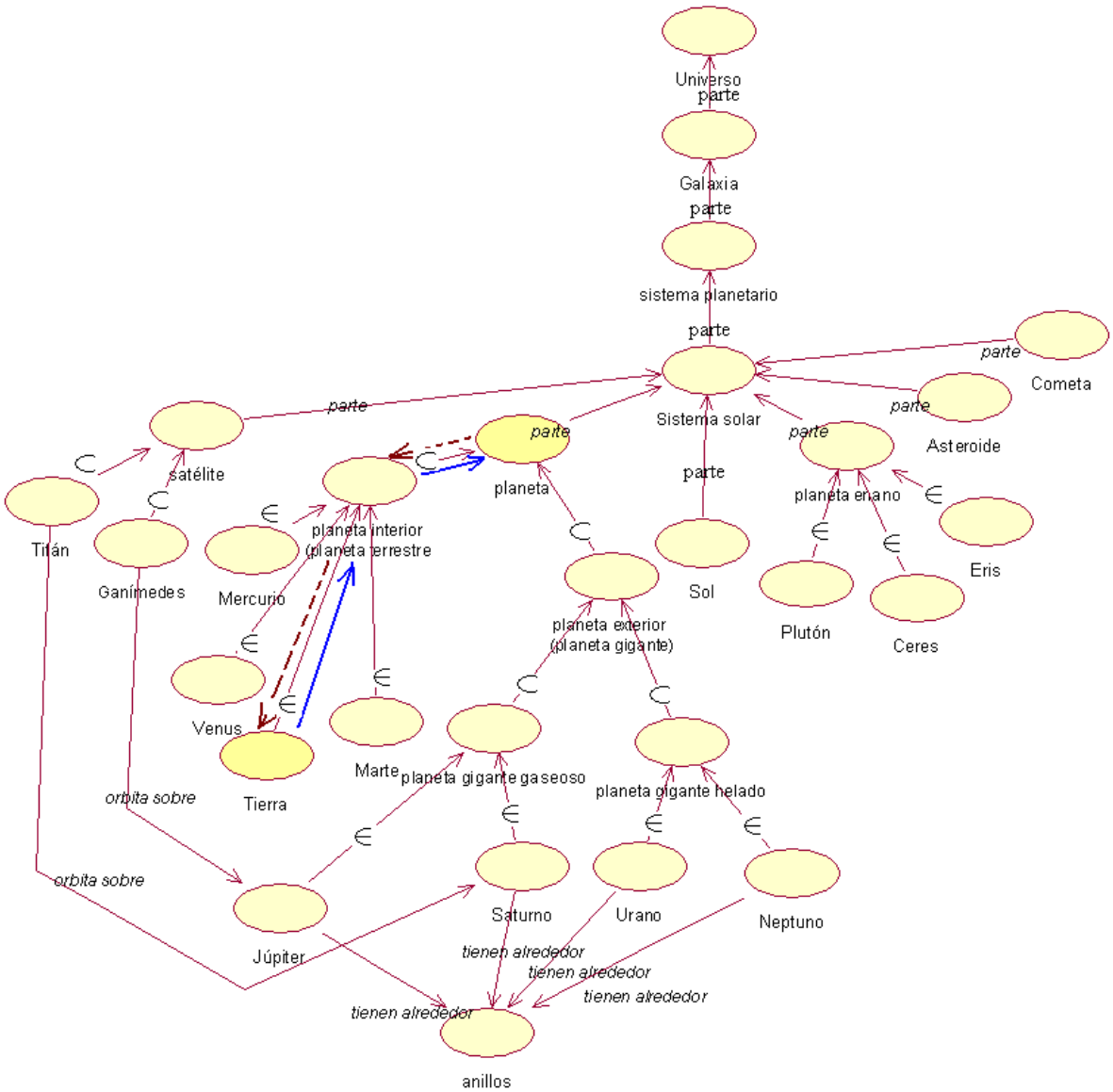

Figura 1. Ontología del sistema solar, donde ∈ significa pertenencia y ⊂ subconjunto. Las líneas azules se explican en la sección 7.2. Ontología hecha a partir del documento http://es.wikipedia.org/wiki/Sistema_Solar.

2.2 En qué consiste la fusión de ontologías (OM)

Considerando dos ontologías A y B, se trata de unirlas en una tercera ontología C (formando una nueva ontología), de manera general se muestra como:

$$C = A \cup \{c_C \mid c_C = ext(r_A, r_B) \ \forall \ c_A \in A, c_B \in B, c_c \in C, r_A, r_B \in R\}$$

La ontología resultante C es la ontología original A añadida de ciertos conceptos y relaciones de B que la función *ext* extrae.

Donde:

c_A es un concepto de la ontología A, r_A son las relaciones de c_A que existen en A, r_B son las relaciones de c_B que existen en B y c_B es el concepto más similar *cms* en B a c_A;

$\cup$ indica una unión de ontologías. Es una unión "cuidadosa" y no la unión de conjuntos.

ext(r_A, r_B) es el algoritmo que complementa las relaciones r_A que ya están en C con aquellas de c_B (que están en B) que no contradicen el conocimiento de A.

Al aplicar *ext* a cada concepto c_A de A, el algoritmo OM extrae de B el conocimiento "adicional" que no estaba presente en A y lo agrega al resultado C. Esta extracción debe hacerse con cuidado, para no introducir inconsistencias, contradicciones o información redundante en C. El algoritmo *ext* es extenso, y viene explicado en las secciones 3.2 a 3.6.

2.3 Conocimiento previo usado por OM

OM se apoya de algunas bases de conocimiento y recursos que le ayudan a detectar contradicciones, encontrar sinónimos, etc. Estas son:

1.- artículos y conectores tales como (en, el, para, este, y, o, etc.) que son ignorados en el nombre de las relaciones.

2.- palabras que cambian o niegan la presencia de conceptos en los nombres de las relaciones, tales como: excepto, sin. Por ejemplo: Amapola sin Peciolo. Significa que el concepto Peciolo no forma parte del concepto Amapola.

3.- Una jerarquía de conceptos. Esta jerarquía es como un árbol de conceptos donde cada nodo es un concepto en otros casos es un conjunto de conceptos de un mismo tipo y en otros es una partición de conceptos (los conceptos que forman la partición son mutuamente exclusivos y juntos forman la partición). La jerarquía representa una taxonomía de términos relacionados entre sí y se usan para medir la confusión (§7.2) y después pueden ser usados para la detección de sinónimos. Más en [Guzmán, A., and Levachkine, S., 2004].

3 Descripción general del método OM para fundir ontologías

De manera general se presenta el algoritmo de fusión OM:

1. Copia la ontología A hacia C.
2. Partiendo del concepto raíz $c_{Raíz}$ en C.
3. Busca en B su concepto más similar c_B (usando COM expuesto en [Cuevas, A., and Guzmán, A., 2005]), el concepto más similar también es conocido como *cms*.
4. Si hay un *cms* en B se adicionan nuevas relaciones, nuevos conceptos, se verifican sinónimos, detectan y resuelven algunas inconsistencias (usando la teoría de la confusión [Guzmán, A., and Levachkine, S., 2004]), se verifica e impide la copia de las relaciones redundantes.
5. Si no hay un *cms* accede al siguiente nodo c_C (hijo de la raíz) y regresa a 3

Si en el paso 4 no se resuelven las inconsistencias se conserva la relación inconsistente en C (la ontología resultante). Más detalles sobre este algoritmo se explican en [Cuevas, A., 2006].

OM se sustenta en dos trabajos importantes:

- El comparador COM [Guzmán, A., and Levachkine, S., 2004], que toma un concepto c_A en una ontología A y halla el concepto más similar c_B en una ontología B.
- La teoría de la confusión (§7.2), que obtiene el grado de confusión de usar un concepto *r* en lugar de otro concepto *s* y la confusión de usar *s* en lugar de *r*.

3.1 Ventajas de OM

OM se desarrolló en el Centro de Investigación en Computación del Instituto Politécnico Nacional (CIC-IPN) en México, es producto de una tesis doctoral, expuesto en: [Cuevas, A. 2006]. Un artículo que la condensa es [Cuevas, A., and Guzman, A., 2007]. Es un algoritmo totalmente automático pues no usa la participación del usuario y es robusto porque se pueden unir ontologías inconsistentes entre sí. En el proceso intervienen dos ontologías A y B para formar una tercera ontología C. Las ventajas de OM se presentan a continuación:

1. Realiza el proceso de manera automática (eliminando las redundancias, resolviendo algunas inconsistencias y los datos imprecisos, por ejemplo: "*Nació en México*" y "*Nació en Guelatao*"). La teoría de la confusión (sección 7.2) ayuda a resolver algunas inconsistencias. Finalmente, si la inconsistencia no puede ser resuelta, OM prefiere el conocimiento de A.
2. Supone que tanto A como B son cada una consistentes (consigo mismas), aunque cierta información en A podría ser inconsistente con otra información en B. Identifica también homónimos como *pico* (herramienta), *pico* (trompa de un ave), y *pico* (cima de una montaña) y no los fusiona como un mismo concepto.

OM abre la posibilidad de que una computadora pueda amasar cada vez más conocimiento sobre un mismo tema, mediante la fusión de ontologías relacionadas (sacadas de distintos documentos). Pero hace falta un *parser* que convierta documentos en ontologías, proceso manual en la actualidad.

3.2 Remoción de relaciones redundantes

En la Figura 2 Norteamérica es subconjunto de Continente y también de América, por tanto se elimina la relación primera (línea punteada), ya que está se encuentra implícita en América como subconjunto de Continente.

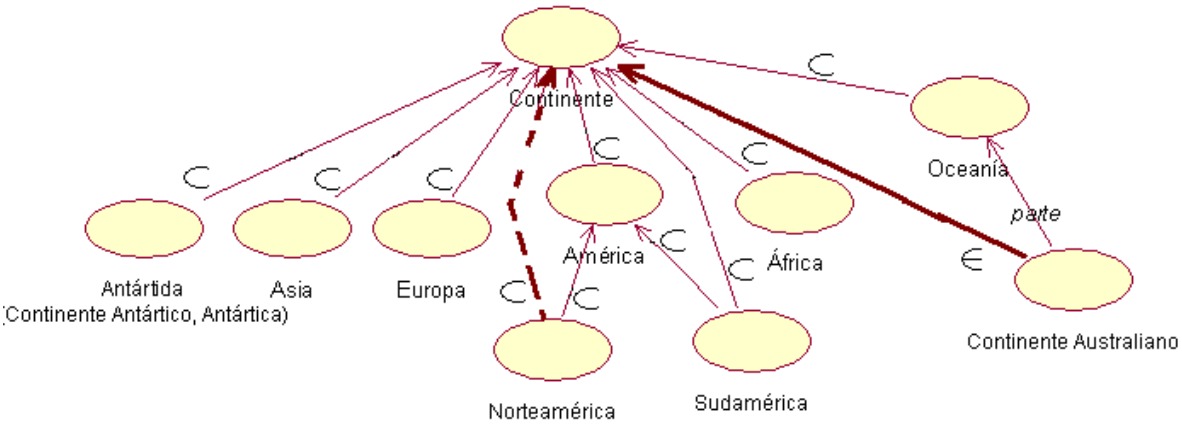

Figura 2. En la ontología el nodo Norteamérica es subconjunto de América y a la vez subconjunto de Continente (redundancia que se elimina), mientras que Continente Australiano es un tipo de Continente y es parte de Oceanía (se conserva). Documento fuente: Wikipedia: http://es.wikipedia.org/wiki/Ant%C3%A1rtida

3.3 Identificación y tratamiento de sinónimos

En la Figura 3 se presenta la ontología A con el concepto Piel como parte del concepto Dinosaurio mientras que en la ontología B se observa el concepto Epidermis como parte del concepto Dinosaurio. OM "descubre" que Epidermis es sinónimo de Piel porque en la definición de este se encuentra la palabra Piel por tanto almacena en C el concepto Piel con las descripciones Piel y Epidermis. Lo mismo sucede con los conceptos Dura y Rígida.

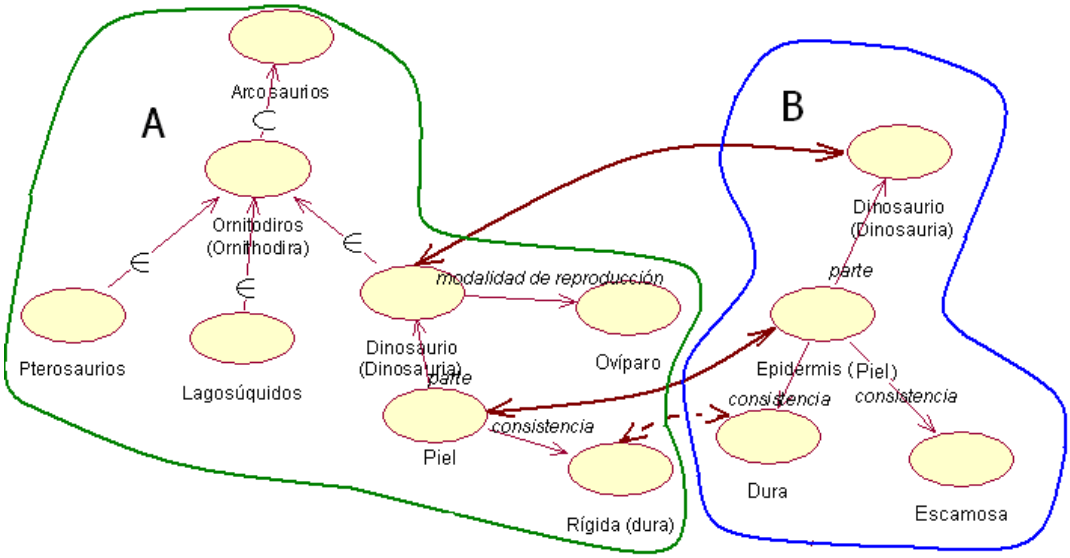

Figura 3. Representación de las ontologías A y B donde se identifican los conceptos Piel y Epidermis como sinónimos, así como Rígida y Dura. Dinosauria es la denominación científica de Dinosaurio (ver http://es.wikipedia.org/wiki/Dinosauria). A y B se muestran parcialmente. Las flechas con dos cabezas indican que tres nodos de A han casado con tres nodos de B (caso "A" de COM, ver sección 7.1)

La Figura 4 representa la ontología C resultante en la cual se puede ver que se ha añadido el nuevo concepto Escamosa una vez reconocido los conceptos Piel y Epidermis como sinónimos.

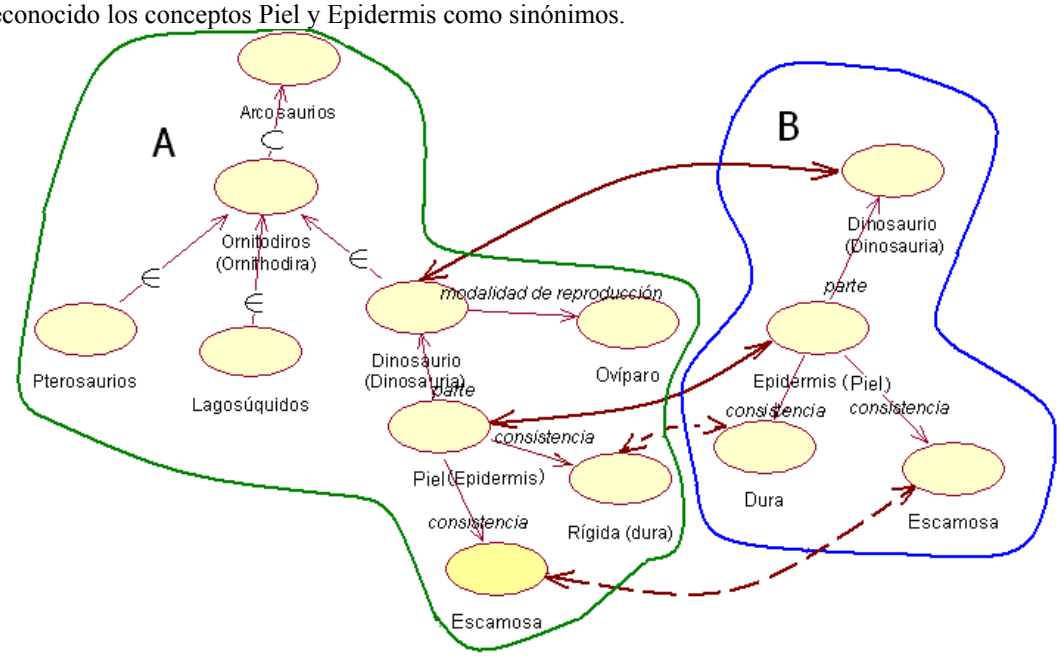

Figura 4. El nodo Epidermis en B es sinónimo de Piel en A, porque comparten palabras en sus definiciones, por lo tanto se añade el concepto Escamosa a Piel en A. Se muestran parcialmente las ontologías A y B. Las líneas con dos cabezas indican casamiento entre conceptos de A y B

Este mismo proceso de identificación de conceptos sinónimos se aplica a la identificación de enlaces o relaciones sinónimas.

3.4 Completando o complementando dos conceptos

El ejemplo representado en la Figura 5 muestra la Ontología A con el concepto Gato como antecesor de Mamífero, igual que en la ontología B, este es un ejemplo del caso A de COM (ver sección 7.1) en la cual coinciden los conceptos y los antecesores. Lo interesante de este ejemplo es que el concepto Silbido no se ha identificado como sinónimo en las características de Gato en A, por tanto se añade como nueva característica en la ontología resultante C.

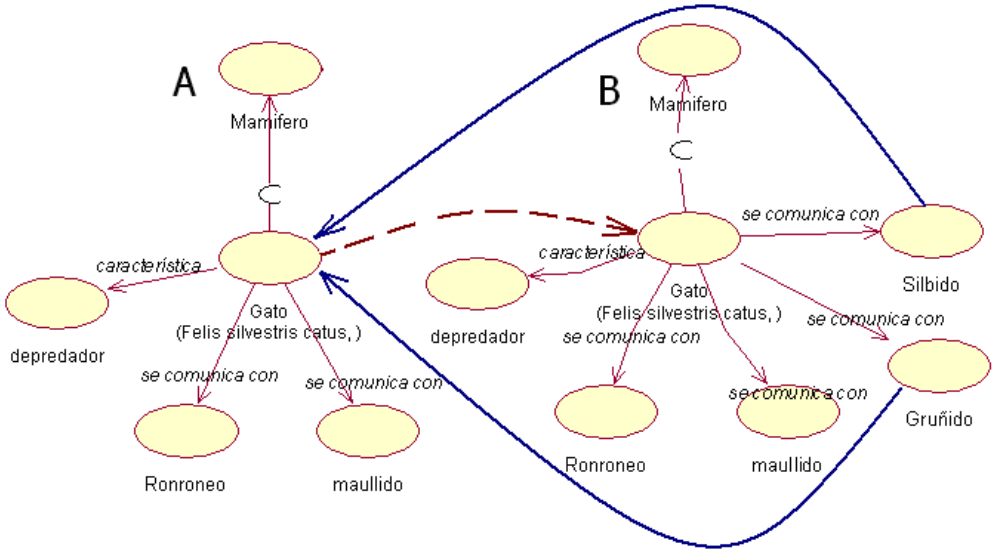

Figura 5. Complementación o enriquecimiento de conceptos: se identifican los conceptos Gato en A y B como iguales y se enriquecen las características de A con las de B

3.5 Conceptos Homónimos

En la Figura 6 se presenta un caso de ambigüedad en el concepto Gato de la ontología A y el concepto Gato en la Ontología B. OM los reconoce como conceptos diferentes porque sus definiciones son diferentes Gato en A significa *Felis silvestres catus* y en B significa *vieja danza criolla alegre y ágil*, sus características también son distintas y sus antecesores también lo son, por lo que añade a la ontología C resultante ambos árboles de conceptos como ramificados separadamente bajo distintos antecesores.

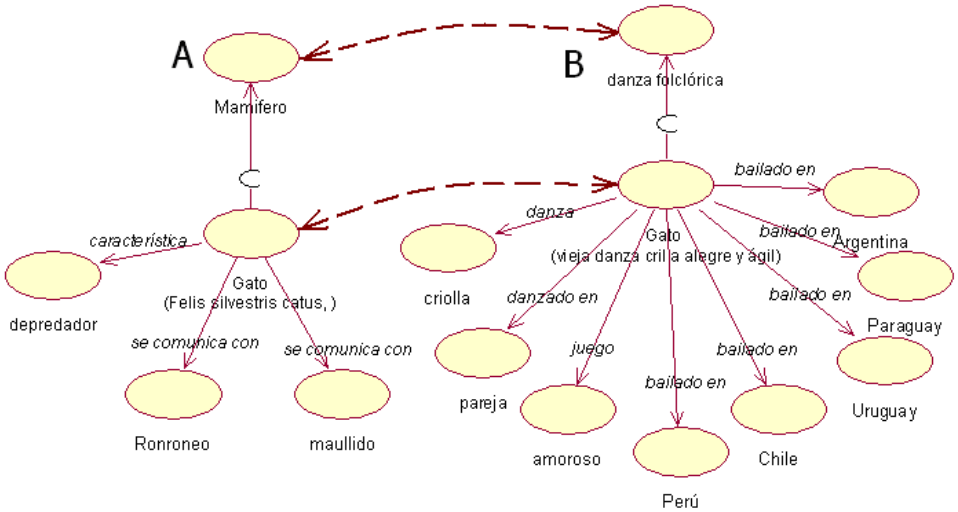

Figura 6. Los conceptos Gato en A y en B comparten la misma sintaxis pero distinta semántica, por lo que se añaden como conceptos diferentes en la ontología C, no mostrada en la figura

3.6 Manejo de particiones y subconjunto

En la Figura 7 se presenta la ontología A en la cual el concepto Supercontinente se identifica en al B con su concepto más similar, particularmente Supercontinente en A tiene dos subconjuntos Gondwana y Lancrasia, mismos que en B son particiones de Supercontinente. OM prefiere registrar en C la partición Desintegración y reconocer los elementos como mutuamente exclusivos y colectivamente exhaustivos.

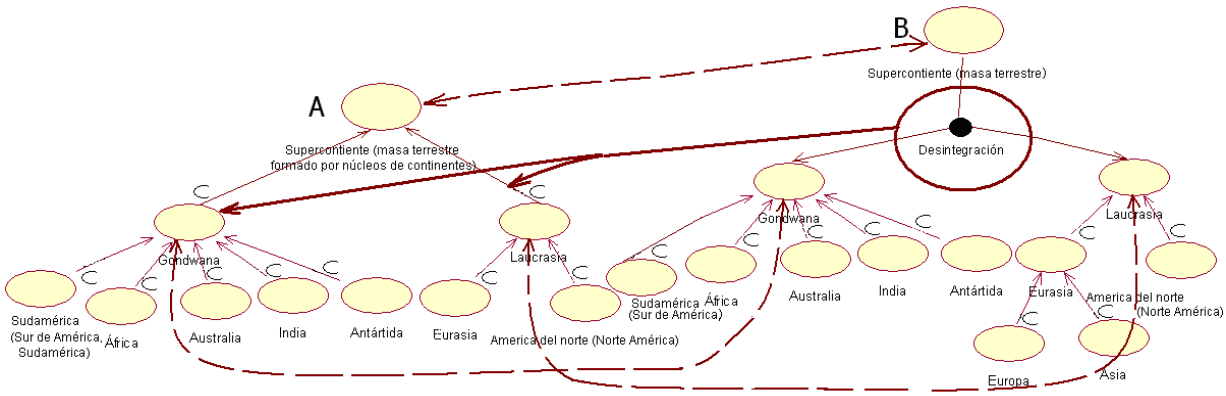

Figura 7. La partición llamada Desintegración del concepto Supercontinente en B se puede añadir a la ontología resultante C, considerando a los conceptos Gondwana y Lancrasia de A como los elementos de esta partición (que también están en B), de tal manera que estos conceptos no solo son subconjuntos de Supercontinente sino conceptos mutuamente exclusivos y ambos colectivamente exhaustivos

La partición registrada en C indica que cada una de las partes de la partición forma el todo y no existen nuevas partes, tampoco falta otras.

Se conserva cada ramificación de los elementos de la partición en B y se añaden nuevos de acuerdo a la respuesta de COM. La Figura 8 presenta la ontología resultante C.

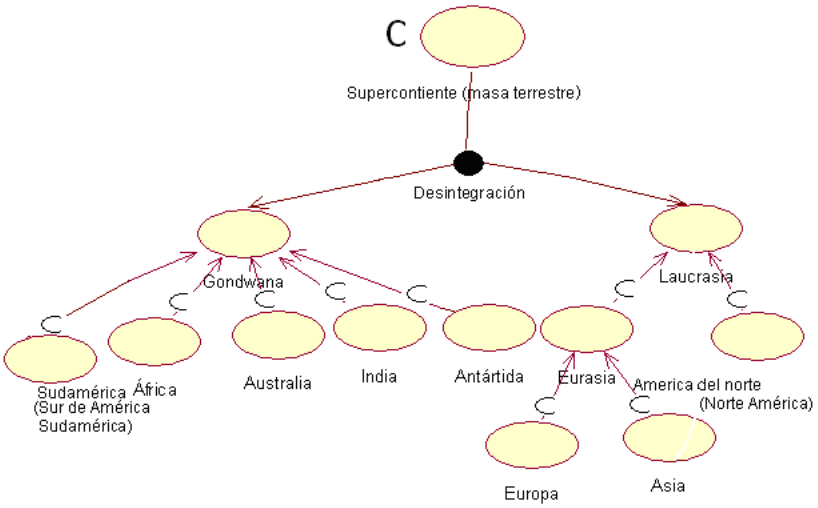

Figura 8. El concepto Supercontinente cuenta con una partición Desintegración con dos elementos que antes eran solo subconjuntos, ahora son más que subconjuntos, son parte de una partición

3.7 Aún se carece de herramientas

En la actualidad casi todas las ontologías que existen disponibles en la Web son taxonomías de un área en especial, son ontologías específicas (ontología de vinos, de virus computacionales, de elementos químicos, etc.) y usan en gran medida la relación "subconjunto." No existe una ontología del conocimiento común.

No existe una herramienta que convierta páginas Web o documentos de texto a ontologías, un algoritmo se encuentra en proceso de desarrollo [Nery, P. 2007]. Por lo anterior, no se aprecia la importancia de la fusión de ontologías.

El algoritmo OM se encuentra en fase de prueba y re-análisis para fortalecer su proceso aprovechando la base de conocimiento de WordNet [Fellbaum, C., 1999] y usarlo para complementar la búsqueda del concepto más similar.

Amén de lo anterior, OM se fortalecerá un contestador de preguntas amplias que está en proceso de construcción [Botello, A., 2007], el que no trabaja sobre una ontología unificada, sino sobre bases de datos heterogéneas.

4 Resultados

Las ontologías fueron obtenidas manualmente de varios documentos descritos en [Cuevas, A., 2006]. Cada par de ontologías a unir describen el mismo tema, por ejemplo, las dos ontologías de tortugas son el resultado de dos documentos hallados en distintos sitios de Internet y cada uno describe tortugas y así, con los otros casos reales. Las ontologías obtenidas fueron unidas (automáticamente) por OM y la validación de la ontología final se hizo de forma manual, obteniéndose buenos resultados.

La primera columna de la tabla 1 presenta las ontologías probadas así como el tiempo que se tardó OM para fusionarlas. La fusión que demoró más es la de *100 Años De Soledad* debido a que tiene más relaciones y OM verifica detalladamente los elementos de cada relación, como se presentó en los algoritmos de este artículo. En la segunda columna se muestra la cantidad de relaciones de A y B que fueron fusionadas. En la columna siguiente se observa la cantidad de conceptos fusionados, el resultado de la fusión manual comparado con la fusión automática de OM en la cual, en algunos casos resultó ser distinta. En la penúltima columna se muestra el error numérico y en la última columna el porcentaje de la eficiencia de OM.

Cálculo del error

El error se ha calculado de la siguiente manera:

$$error = \frac{númeroDerelacionesyConceptosMalcopiadosEnC}{númeroTotalDe\,\mathrm{Re}lacionesyConceptosCopiadosManualmenteEnC}$$

Cálculo de la eficiencia

La eficiencia se ha calculado de la siguiente manera:

$$eficiencia = 100 * \frac{númeroDe\,\mathrm{Re}lacionesyConceptosCopiadosCorrectamenteEnC}{númeroTotalDe\,\mathrm{Re}lacionesyConceptosCopiadosManualmenteEnC}$$

Tabla 1. Funcionamiento de OM en algunos ejemplos reales

Ontologías	Relaciones	Conceptos	error	% Efic.
Sistema solar §2.1 (4 seg.)	Las 45 relaciones de B se añadieron y fusionaron correctamente a las 56 de A, dando un total de 59 relaciones en C	Los 60 conceptos de B[1] se añadieron y fusionaron correctamente a los 79 de A[2], dando como resultado 75 conceptos en C	0	100
Neurotransmisor y Esquizofrenia[3] (2 seg.)	Las 79 relaciones de Neurotransmisor (A) se añadieron y fusionaron con las 51 de Esquizofrenia (B) dando un total de 127 relaciones en C, el método manual dio 129 (2 de 129 no fueron copiados)	Los 56 conceptos de Neurotransmisor (A) se añadieron y fusionaron con los 26 de Esquizofrenia dando un total de 77 nodos, el método manual dio 79 (faltó 2 de 79 conceptos)	0.019	98
Continente §3.2 (3 seg.)	Las 40 relaciones de B se añadieron y fusionaron correctamente a las 34 de A, dando un total de 46 relaciones en C	Los 54 conceptos de B[4] se añadieron y fusionaron correctamente a los 50 de A[5], dando como resultado 66 conceptos en C	0	100
Ontologías Inconsistentes (1 seg.)	Las 3 relaciones de B se añadieron y fusionaron con las 4 de A dando un total de 7, hubieron 2 inconsistencias (2 relaciones que no se definieron, una por cada ontología), C tuvo 1 inconsistencia, OM resolvió otra.	Los 5 conceptos de B se añadieron y fusionaron con los 6 conceptos de A, dando un total de 9 conceptos en C, hubieron 5 inconsistencias (3 conceptos en A mal clasificados y 2 en B), C tuvo las mismas 5 inconsistencias	0	100
Dinosaurio (3 seg.)	Las 40 relaciones de B fueron añadidas y fusionadas correctamente a las 41 de A, dando un total de 47 relaciones en C	Los 45 conceptos de B[6] se añadieron y fusionaron correctamente a los 44 de A[7], dando como resultado 53 conceptos en C	0	100
100 Años de Soledad (10 minutos)	Las 283 relaciones de B se añadieron y fusionaron con las 231 de A dando un total de 420, el método manual dio 432 (12 de 432 no fueron copiados)	Los 126 conceptos de B se añadieron y fusionaron con los 90 de A dando un total de 141, el método manual dio 149 (8 de 149 no fueron copiados)	0.034	96.5
Oaxaca (5 min.)	Las 43 relaciones de B fueron añadidas y fusionadas a las 61 de A dando un total de 96 relaciones en C	Los 117 conceptos de B se añadieron y fusionaron con los 234 de A dando un total de 309, el método manual dio 310 (faltó 1 de 310 conceptos)	0.002	99.7

[1] La ontología B se ha extraído del documento: http://www.solarviews.com/span/solarsys.htm

[2] La ontología A se ha extraído del documento: http://es.wikipedia.org/wiki/Sistema_Solar

[3] Se agradece a Paola Nery Ortiz www.geocities.com/paolaneriortiz/ por su apoyo en la conversión manual dos documentos a ontologías: el primero acerca del neurotransmisor es.wikipedia.org/wiki/Neurotransmisor y el segundo acerca de la esquizofrenia www.nimh.nih.gov/publicat/spSchizoph3517.cfm , estas ontologías se han fusionado automáticamente usando OM y comprobado el resultado de forma manual.

[4] La ontología B surgió del documento: http://es.wikipedia.org/wiki/Redefinici%C3%B3n_de_planeta_de_2006

[5] La ontología A se ha extraído del siguiente documento: http://es.wikipedia.org/wiki/Continente

[6] La ontología B surgió del documento: http://www.dinosaurios.net/

[7] La ontología A surgió del documento: http://es.wikipedia.org/wiki/Dinosauria

5 Discusión

5.1 Sobre la ontología del "sentido común"

Aún cuando pudiese construirse un OM mejorado (OM*) o software similar que fuese amasando conocimiento, cabe señalar que hay cierto conocimiento que, por no estar en ningún documento, no puede formar parte del repositorio de información de este. Refiérase al "conocimiento común", o "del sentido común" (que era lo que CYC [Reed, S. L., and Lenat, D., 2002] iba a captar en su ontología), por ejemplo: *cuando llueve*, *llueve agua*, *llueven gotitas redondas*, *llueve de arriba para abajo*. Otro ejemplo: *Los padres siempre son más viejos que los hijos*. *Cuando entierran a alguien*, *normalmente lo entierran con todos sus órganos y partes de su cuerpo*. Otro más: *De las uvas se hace el vino, pero del vino no se hacen uvas*.

Esta información se adquiere por las personas, por el hecho de vivir e interactuar con el medio ambiente, de observar la naturaleza y la vida cotidiana. Y no está escrita en documentos, pues serían poco leídos por personas, que "ya saben todo eso".

OM* no puede procesar estos documentos, puesto que no existen. De momento esto no es motivo de preocupación, se puede avanzar bastante en amasar información en documentos existentes, y adquirir el conocimiento común en otra forma o en otra etapa.

5.2 Análisis sintáctico versus análisis semántico

Actualmente, los métodos de fusión de ontologías, exceptuando a HCONE-Merge [Kotis, K., and Vouros, G., Stergiou, K., 2006] realizan la alineación y fusión comparando las etiquetas o nombres de los conceptos así como sus vecindades, HCONE-Merge usa la base de datos de WordNet [Fellbaum, C., 1999] para encontrar el significado de los conceptos en la alineación, luego OM usa la definición de los conceptos, sus vecindades y sus características, estas últimas se verifican mediante un proceso recursivo ya que cada característica puede ser también un concepto y se analiza la definición de este, sus vecindades y sus características y así sucesivamente. Este proceso de búsqueda recursiva se puede interpretar como un proceso de búsqueda semántica en los conceptos llamados también análisis semántico.

El análisis semántico difiere del análisis sintáctico en el sentido de que en él se verifica el significado del concepto, sus relaciones con otros conceptos y sus características, a lo que el análisis sintáctico verifica las palabras del concepto y sus relaciones pero no indaga más acerca de las características de los conceptos anidados en él, tampoco verifica que cada característica puede ser también un concepto y por tanto puede tener otras relaciones.

El análisis semántico tiene más posibilidades de encontrarle más significado al concepto, mientras que en el sintáctico se reducen estas posibilidades.

5.2 Comprobando automáticamente la fusión

Hoy en día, la fusión se comprueba manualmente (sección 1.4). Se podría comprobar más automáticamente usando el contestador de preguntas que se construye en [Botello, A., 2007]. Como éste está diseñado para contestar preguntas sobre el resultado de la fusión de base de datos heterogéneas, habrá que adaptarlo para analizar ontologías.

5.3 Otras herramientas para el fusionador OM

OM se podría completar (y convertirse así en OM*) con un par de herramientas adicionales:

- el parser o convertidor de documentos de texto a ontologías [Nery, P., 2007]. Puede verse como un "preprocesador" para OM. Nos brindará ontologías a fusionar por OM;
- el contestador de preguntas complejas, mencionado en §1.4 y en §5.2, que nos permitirá explotar o utilizar para fines prácticos el conocimiento que OM* amasa.

Además, es factible enriquecer OM mediante adiciones que mejoren la fusión. A continuación se citan algunas:

1.- Recursos linguísticos a accesar: WordNet, WordMenu, etc. Con la finalidad de:

a. Desambiguar, es decir, asignar a una palabra un concepto o "sentido" o acepción. Ejemplo: desambiguar palabras, por ejemplo pico. De la siguiente manera: (i) usando glosas extendidas como la que se expone en [BannerJee, S., and Pedersen T., 2003]; (ii) usando WordNet como ontología puente (virtual, quizá) [Kotis, K., and Vouros, G., Stergiou, K., 2006].

b. Para conocer más acerca de la relación "part" (parte de), es decir, qué concepto es parte de qué otro concepto, ejemplo, un barco se compone de proa, popa, estribor, babor, cubierta, propela… (usando WordMenu).

2. Reglas para manejar eventos que transcurren en el tiempo (procesos). Aquí surgen ciertas restricciones:
   a. Algunos eventos no dicen cuándo sucedieron, por ejemplo: "*abandonó el Seminario Santa Cruz*",
   b. Otros eventos se indican aproximadamente, por ejemplo: "*recorrió gran parte de la República mexicana con los archivos de la nación después de ser destituido de su cargo*" o en forma relativa.
   c. Otros son del tipo "siempre", por ejemplo: "*Rosa es la hermana de Benito*", "*Benito es Oaxaqueño*". Estos son considerados actualmente en OM.

   Esto origina que un evento tenga dos fechas (de inicio y fin) que sean conocidas (constantes) o no sean conocidas completamente (variables), por ejemplo: "*abandonó el Seminario Santa Cruz en* ***t***", otro ejemplo: "*ingresó a la universidad para estudiar Derecho en el momento* ***u***" donde ***t*** y ***u*** son desconocidos porque pueden indicar tiempo o espacio. Además pueden haber restricciones sobre esas variables (***t*** > 1812, ***t*** > ***u***). Existen trabajos que han presentado avances sobre este tema [Puscasu, G., Ramirez Barco P, 2006].
3. Ídem para manejar eventos que ocurren en el espacio (localización en el espacio). Puede ser relevante la ontología geoespacial [Montes de Oca, V., Torres, M., Levachkine, S. and Moreno, M., 2006].

6 Conclusiones

Con la llegada de OM, es posible ahora fusionar dos ontologías de manera automática, sin intervención humana. Las pruebas efectuadas muestran calidad en los resultados obtenidos por OM. OM detecta inconsistencias y resuelve algunas, detecta sinónimos, homónimos, información redundante y con diferente grado de detalle o precisión.
Hacen falta un parser que convierta documentos en lenguaje natural a ontologías, y un razonador que responda (usando la ontología resultado de OM) preguntas complejas.

7 Apéndice A Trabajos en los que se sustenta OM

El algoritmo COM [Guzmán, A., and Levachkine, S., 2004] toma un concepto C_A en una ontología A y halla el concepto más similar C_B en una ontología B (§7.1).

El algoritmo de la teoría de la confusión (§7.2) obtiene el grado de la confusión de usar un concepto *r* en lugar de otro *s* y la confusión de usar *s* en lugar de *r*.

7.1 El comparador de Ontologías Mixtas COM

Para hallar la similitud o semejanza entre los nodos de una ontología a otra se usa el algoritmo Comparador de Ontologías Mixtas COM [Guzmán, A., and Olivares, J., 2004] que usa las definiciones (palabras) y las relaciones de tales conceptos usando 4 casos distintos en el proceso.

1.- Caso A: el concepto de C_A casa con C_B en B y los antecesores P_A y P_B también
2.- P_A casa con P_B pero no C_A con C_B
3.- C_A coincide con C_B pero no P_A con P_B
4.- No coinciden C_A con C_B, tampoco P_A con P_B

7.1.1 Caso A: C_A casa con C_B y P_A casa con P_B

En B se buscan dos conceptos C_B y P_B, de manera que la definición de P_B coincida con la mayoría de las palabras que definen P_A y la mayoría de las palabras que definen C_B coincidan en la definición de C_A el algoritmo devuelve:

* El C_B (conocido también como *cms*) en B,
* El valor *vs* (*valor de similitud*) = entre 0 y 1.

Los sub casos son los siguientes:

1) *Caso A Papá*: C_A casa con C_B y P_A con P_B.
2) *Caso A Abuelo*: C_A casa con C_B pero P_A con A_B, el abuelo de C_B.
3) *Caso A Abuelo simétrico*: C_A casa con C_B pero A_A, el abuelo de C_A con P_B.
4) *Caso A Bisabuelo*: C_A casa con C_B pero P_A con B_B, el bisabuelo de C_B.
5) *Caso A Bisabuelo simétrico*: C_A casa con C_B pero B_A, el bisabuelo de C_A con P_B.

Este caso considera el hallar conceptos entre ontologías definidas bajo el mismo tema y estructura, es decir, ontologías parecidas o igualmente definidas. Por ejemplo en la ontología A existe un concepto Herramienta con el concepto Martillo como sucesor y en B también.

7.1.2 Caso B: Los papás P_A y P_B coinciden, pero no hay C_B

Se encuentra P_B pero no C_B, se llama recursivamente a COM con P_A como parámetro para confirmar que P_B es antecesor de C_A. Si el P_B primo o candidato (P_B') hallado es la raíz de la ontología ($O_{BRaíz}$) el algoritmo concluye sin éxito; si no, se busca en B para cada hijo de P_B, aquél, cuya mayoría de propiedades y valores, coincidan (llamada recursiva a COM) con las correspondientes de C_A. Es decir, se busca en B al hijo de P_B que tenga la mayor cantidad de las propiedades (y valores) de C_A. Si el candidato C_B' tiene además hijos, se ve si éstos coinciden (usando llamadas recursivas a COM) con los hijos de C_A. Si se encuentra un C_B' con las propiedades deseadas, el algoritmo concluye con éxito indicando el C_B' encontrado.

En otro caso, se intenta hallar el C_B' entre los hijos del padre (en B) de P_B, es decir, entre los hermanos de P_B; en caso necesario, entre los hijos de los hijos de P_B, o sea, entre los nietos de P_B Si no se encontró un C_B', entonces el más cercano a C_A es P_B, por lo que COM devuelve el mensaje "hijo de P_B " (significa que un hijo de P_B que aún no existe en B es el más similar a C_A) y el algoritmo concluye.

Este caso considera la posibilidad de que se existan definiciones nuevas de conceptos que son subconjuntos o tipos de un concepto general o antecesor, por ejemplo en A existe un concepto Muebles y en B también pero los tipos de Muebles de A son distintos a los de B.

7.1.3 Caso C: coincide C_A con C_B pero no hay P_B

Si se halla C_B pero no P_B, se busca si el abuelo en B de C_B es similar a P_A o si el bisabuelo de C_B en B es similar a P_A (esto ya fue comentado en el Caso A). Si se halla, entonces el concepto en B más similar a P_A es el abuelo o el bisabuelo de C_B y concluye el algoritmo. Si no se halla, se verifica si la mayoría de las propiedades (y sus valores correspondientes) de C_A coinciden con las de C_B y si la mayoría de los hijos de C_A coinciden (usando *hallar*) con la mayoría de los hijos de C_B; si las propiedades y los hijos coinciden, entonces la respuesta es C_B y concluye el algoritmo aunque no se haya hallado en B al P_B que corresponda al concepto P_A en A. Si solamente una parte de propiedades e hijos son similares entonces la respuesta es "probablemente C_B" y concluye el algoritmo. Si ninguna propiedad ni hijos son similares la respuesta es "no existe" y concluye el algoritmo.

Este caso considera los conceptos que comparten la misma etiqueta, definición o léxico pero con distinta semántica, por ejemplo: en la ontología A hay un concepto Radio con el antecesor Medio de comunicación y en la ontología B un concepto Radio con el antecesor Geometría.

7.1.4 Caso D: No hay C_B ni P_B

Si no se encuentra C_B ni P_B, entonces la respuesta es "no existe" y concluye el algoritmo. Esta situación es común cuando se tratan de dos ontologías totalmente diferentes, por ejemplo: una ontología con temas de "Sistemas de Información" con otra ontología con temas de "Recursos Naturales".

7.2 Confusión

La confusión, contradicción o inconsistencia surge cuando una relación en C_A que hace incompatible, contradice o hace inconsistente a otra relación en C_B. Por ejemplo la relación: (forma; Tierra; redonda) en A y (forma; Tierra; plana) en B. Nótese que la contradicción surge de *dos relaciones* explícitas (estas relaciones se detallan en [Cuevas, A. 2006]). Es decir, una contradicción surge porque A le da una semántica (expresada en una relación) a un concepto (Tierra, digamos) y B le da otro significado.

Sean *r* y *s* dos valores (dos nodos) en una jerarquía con altura[8] *h* y sea *r*' cualquier ascendiente de *r*.

La función CONF(*r*, *s*) llamada *confusión absoluta* que resulta de usar *r* en vez de *s* (el valor deseado) es:
CONF(*r*, *r*) = CONF(*r*, *s*) = 0 cuando *s* es algún ascendiente de *r*;
CONF(*r*, *s*) = 1 + CONF (*r*, *padre_de*(*s*)) en otro caso.
CONF(*r*, *s*) puede variar meramente mediante la inserción de nodos adicionales en el camino descendente entre *r* y *s.* Para evitar esto, definiremos la confusión relativa, conf, como sigue:

Definición
La función conf(r, s) que resulta de usar *r* en vez de *s* es:

$$conf(r,s) = \frac{CONF(r,s)}{h}$$

La función conf(*r*, *s*) es la confusión absoluta CONF(*r*, *s*) dividida por *h,* la altura de la jerarquía.
conf halla la confusión de usar un concepto *r* en lugar de otro s(hay más detalles de su funcionamiento en [Guzmán, A., and Levachkine, S., 2004]); no es una función simétrica.
La función *conf*(*r*, *s*) parte de dos conceptos (valores) *r* y *s* que están en una jerarquía de conceptos J. Se obtiene el valor de usar *r* en lugar de *s* y el valor de usar *s* en lugar de *r*.
Se ubica en la posición del concepto *r* en la jerarquía J trasladándose hasta el concepto *s* de misma jerarquía, sumando solo los niveles descendentes (si los hay), de la ruta hacia *s*, luego se divide esta suma entre la altura del árbol (la jerarquía) a este resultado se le conoce como valor de la confusión *vc*, luego parte de la posición *s* de la jerarquía J trazando una ruta hacia el concepto *r* sumando los niveles descendentes y dividiendo esta suma entre la altura del árbol (su *vc*).
Ejemplo. Referirse a la ontología de la figura 1 considerándola como si fuera una jerarquía. conf(Tierra, Planeta) = 0, mientras conf(Planeta, Tierra)=2. (líneas azules en la figura 1).
Ejemplo: si en una jerarquía se tiene que (*parte de,* San Pablo Guelatao, Sierra de Ixtlán), (*parte de,* Sierra de Ixtlán, Oaxaca); (*parte de,* Oaxaca, México), entonces CONF (San Pablo Guelatao, México) = 0; CONF (México, San Pablo Guelatao) = 2, y OM toma el valor mínimo de ellos (cero), por lo que deduce que San Pablo Guelatao es más específico que México, y frente a las afirmaciones de la ontología A = (*nació en,* Benito Juárez, San Pablo Guelatao) y B = (*nació en,* Benito Juárez, México), detecta que no hay contradicción y almacena en C (*nació en,* Benito Juárez, San Pablo Guelatao). Para resolver este tipo de "contradicción", OM recorre ambas rutas y teniendo ambos valores de la confusión *vc*, se elige el menor *vc* obtenido de las rutas de *r* hacia *s* y de *s* hacia *r*.

Referencias

BanerJee, S., and **Pedersen T**. "Extended Gloss Overlaps as Measure of Semantic Relatedness". In Proc. of *IJCA-03*, pp. 805-810. México. 2003
Botello, A. "Infiriendo Relaciones Entre Bases de Datos Autónomas". CIC-IPN. Tesis Doctoral en proceso de desarrollo. [2007].
Cuevas, A., and **Guzmán, A**. "Improving the Search for the Most Similar Concept in other Ontology". In proc. *XVIII Congreso Nacional y IV Congreso Internacional de Informática y Computación*. Torreón Coah. México. Octubre 2005.
Cuevas, A., and **Guzman, A**. "A Language and Algorithm for Automatic Merging of Ontologies" a chapter of the book "Handbook of Ontologies for Business Interactions", Peter Rittgen, ed. Idea Group Inc, Publishers. Hershey, PA, USA. *In press*. 2007

[8] La altura de un árbol es el número de aristas que hay en la trayectoria de su raíz a la hoja más distante. Ejemplo: la altura del árbol en la Figura 1 es 8.

Cuevas, A. "Unión de ontologías usando propiedades semánticas". Tesis doctoral. CIC-IPN. México, Diciembre 2006. Disponible en:
http://148.204.20.100:8080/bibliodigital/ShowObject.jsp?idobject=34274&idrepositorio=2&type=recipiente
Dou, D., McDermott, D., and **Qi, P**. "Ontology Translation by Ontology Merging and Automated Reasoning". In Proc. *EKAW Workshop on Ontologies for Multi-Agent Systems*. 2002. -
Fellbaum, C. "WordNet An Electronic Lexical Database". Library of Congress Cataloging in Publication Data.1999.
Fridman, N., and **Musen, M**. "PROMPT: Algoritm and Tool for Automated Ontology Merging and Alignment". In Proc. *Seventeenth National Conference on Artificial Intelligence*. pp 450-455, Austin, TX, USA, 2000.
Gruber, T. "Toward principles for the design of ontologies used knowledge sharing". Originally in N. Guarino & R. Poli, (Eds.), *International Workshop onf Formal Ontology*, Padova, Italy. 1993.
Guzmán, A., and **Levachkine, S**. "Hierarchies Measuring Qualitative Variables". *Lecture Notes in Computer Science* LNCS 2945 [Computational Linguistics and Intelligent Text Processing], Springer-Verlag. 262-274. ISSN 0372-9743. 2004.
Guzmán, A., and **Olivares, J**. "Finding the Most Similar Concepts in two Different Ontologies". *Lecture Notes in Artificial Intelligence* LNAI 2972, Springer-Verlag. 129-138. ISSN 0302-9743. 2004.
Kalfoglou, Y., and **Schorlemmer, M**. "Information-Flow-based Ontology Mapping". Proceedings of the 1st *International Conference on Ontologies, Databases and Application of Semantics* (ODBASE'02), Irvine, CA, USA. 2002.
Kotis, K., and **Vouros, G.**, **Stergiou, K**. "Towards Automatic of Domain Ontologies: The HCONE-merge approach". *Elsevier's Journal of Web Semantic (JWS),* vol. 4:1, pp 60-79. Available on line ant (ScienceDirect): http://authors.elsevier.com/sd/article/S1570826805000259 2006.
McGuinness, D., Fikes, R., Rice, J., and **Wilder, S**. "The Chimaera Ontology Environment Knowledge". In Proceedings of the *Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000)*. Darmstadt, Germany. 2000.
Montes de Oca, V., Torres, M., Levachkine, S. and **Moreno, M**. "Spatial Data Description by Means of Knowledge-Based System". *Lecture Notes in Computer Science*, Vol. 4225, Springer-Verlag (2006).
Nery, P. "Parser para la conversión de documentos de texto a ontologías". Tesis en construcción. CIC-IPN. México. 2007.
Puscasu, G., Ramirez Barco P, *et al*. On the identification of temporal clauses. *LNAI 4293,* 911-921 (MICAI 06). 2006.
Reed, S. L., and **Lenat, D.** "Mapping Ontologies into Cyc". In proceeding of *AAAI Workshop on Ontologies and the Semantic Web*, Edmonton, Canada. 2002.
Stumme, G., Maedche, A. "Ontology Merging for Federated ontologies on the semantic web". In: E. Franconi, K. Barker, D. Calvanese (Eds.): Proc. *Intl. Workshop on Foundations of Models for Information Integration (FMII'01)*, Viterbo, Italy, 2001. lNAI, Springer 2002 (in press).

Referencias de Internet.

1. Loom http://www.isi.edu/isd/LOOM/LOOM-HOME.html
2. wikipedia http://es.wikipedia.org/wiki/Ontolog%C3%ADa