

Integración de Ontologías Datalog \pm Mediante Consolidación a Través de Debilitamiento

Cristian D. Pacifico

**Cristhian A. D. Deagustini
Guillermo R. Simari**

Marcelo A. Falappa

Area de Agentes y Sistemas Inteligentes

Facultad de Ciencias de la Administración - Universidad Nacional de Entre Ríos

Tavella 1424, (3200) Concordia, Entre Ríos - (0345) 423-1400

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial

Instituto de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur - CONICET

Alem 1253 - Bahía Blanca - Buenos Aires - Argentina - (0291) 459-5135

cripac@fcad.uner.edu.ar, (cadd,mfalappa,grs)@cs.uns.edu.ar

Resumen

En sistemas que manejan e intercambian grandes volúmenes de información es necesario contar con métodos adecuados para manejar conflictos que aparecen en la medida que sus bases de conocimiento evolucionan y se integra con otras fuentes de información. Este conocimiento es frecuentemente expresado por ontologías; no solo por el valor expresivo, sino también por las diferentes restricciones semánticas que se pueden aplicar al conocimiento, originadas por el dominio de aplicación o por los propios usuarios.

Esta línea de I+D profundiza en la resolución automatizada de conflictos en ontologías expresadas en Datalog \pm ; ya sea estos generados por inconsistencia o incoherencia. Se pretende definir métodos que resuelven los conflictos que surgen al integrar dos o más ontologías Datalog \pm . Una operación de integración sobre varias ontologías, debe restaurar la consistencia y coherencia en el todo, no solo considerando las características intrínseca de la información integrada, sino también meta-información relacionada con la confiabilidad de cada fuentes, la ocurrencia de información, etc. En particular, se definirán mecanismos resolución de incoherencias e inconsistencias que produzcan alguna forma de debilitamiento o modificación de las reglas; para lograr la mínima pérdida de información en el proceso de integración. Finalmente, se analizará los mecanismos descriptos para extenderlos a la integración con otras fuentes de datos, en particular con Bases de Datos Relacionales.

Palabras Clave: Integración de Bases de Conocimiento, Revisión de Creencias, Representación de Conocimiento, Razonamiento.

1. Contexto

Esta línea de investigación se lleva a cabo en el marco de los siguientes proyectos de investigación:

- **“Representación de conocimiento y razonamiento argumentativo: Herramientas inteligentes para la web y las bases de conocimiento”**. Director: Guillermo R. Simari. 01/01/15 – 31/12/2018. Unidad coordinadora: Universidad Nacional del Sur.
- **“Combinación de Revisión de Creencias y Argumentación para mejorar las capacidades de Razonamiento y modelado de la Dinámica de Conocimiento en Sistemas Multi-agente”**. Director: Marcelo A. Falappa. PIP 112-20110101000. Unidad coordinadora: Consejo Nacional de Investigaciones Científicas y Técnicas.

La línea de I+D se enmarca dentro del ámbito de colaboración entre el Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA) del Instituto de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur/CONICET y el Área de Agentes y Sistemas Inteligentes de la Facultad de Ciencias de la Administración, Universidad Nacional de Entre Ríos.

2. Introducción

2.1. Conflictos en Ontologías

En sistemas que manejan e intercambian grandes volúmenes de información es necesario contar con métodos automáticos y adecuados para el manejo de conflictos que aparecen naturalmente en la medida que el conocimiento evoluciona.

En los últimos tiempos, las Bases de Conocimiento expresadas mediante ontologías ha sido un temática de interés para áreas de investigación y desarrollo. En particular las ontologías son importantes no sólo por el conocimiento que pueden expresar, sino también por las diferentes restricciones semánticas que se pueden aplicar a dicho conocimiento.

El manejo de información conflictiva en bases de conocimiento es un importante problema que debe ser atacado [BQLO7], especialmente cuando se integra conocimiento proveniente de diferentes fuentes [AK05], o cuando tal conocimiento será explotado por procesos de razonamiento automáticos.

El más conocido de conflictos en información es la *inconsistencia*. El concepto se refiere a teorías tales que es imposible encontrarles un *modelo*; es decir, al menos una *interpretación* que haga verdaderas a todas sus fórmulas. En términos sintácticos, un conjunto de fórmulas es consistente si y sólo si para toda fórmula A , no es posible deducir tanto A como $\neg A$.

Adicionalmente, en entornos ontológicos se encuentra otro tipo de conflicto, relacionado a un fenómeno conectado con la inconsistencia pero a su vez con sus características propias: la *incoherencia* [FHP⁺06, QH07]. Este fenómeno surge cuando una porción de conocimiento expresado por un conjunto de reglas no pueda ser aplicado sin generar problemas de consistencia; es decir, el conjunto de reglas no puede ser aplicadas sin violar inevitablemente alguna de las restricciones impuestas al conocimiento, haciéndolas por lo tanto insatisfacibles.

En estos entornos, una posibilidad para tratar estos conflictos y recuperar la coherencia y la consistencia implica modificar la información contenida en la ontología inconsistente de forma tal que la consistencia sea recuperada. De esta forma, se modifica la base de conocimiento original (e inconsistente), de manera que la base de conocimiento modificada, pueda aplicar relaciones de inferencia clásicas de manera segura. Este es uno de los objetivos perseguidos por la teoría de *Revisión de Creencias* [Gäro3, AGM85, Han94].

Siguiendo este enfoque, trabajos previos [Dea16, DMFS16] han presentado de manera teórica, una alternativa para conseguir la consolidación de ontologías *Datalog*[±]. *Datalog*[±] es un framework para realizar consultas de ontologías tratables y otras múltiples aplicaciones, basado en el lenguaje *Datalog* [CGL12], incorporando características tales como cuantificadores existenciales para la cabeza de reglas y restricciones en sintaxis para alcanzar decibilidad y, en una gran parte de los casos, incluso tratabilidad. En particular, la representación de conocimiento en ontologías *Datalog*[±] se lleva a cabo mediante el uso de (a) una *Base de Datos*: un conjunto de átomos que representan hechos acerca del mundo, e. g., *alumno(pedro)* (b) *Tuple-generating Dependencies (TGDs)*: reglas que permiten obtener nuevos átomos mediante la activación de las mismas como ser: *alumno(X) → persona(X)*,

(c) *Equality-generating Dependencies (EGDs)*: reglas que restringen la generación de átomos, por ejemplo: *doctor(D, P) ∧ doctor(D', P) → D = D'*; y (d) *Negative Constraints (NCs)*: reglas que expresan relaciones que no pueden existir entre átomos, e. g.: *alto(X) ∧ bajo(X) → ⊥*.

Un importante beneficio por utilizar bases de conocimientos es la posibilidad de manejar dinámica de conocimiento. Es decir, aceptar nueva información manteniendo la integridad y consistencia de la base. En este sentido, el desafío es eliminar los conflictos que se producen en el proceso de actualización de la base de conocimiento ante las novedades. Para esto, se tomará como marco teórico los operadores de consolidación de ontologías *Datalog*[±] presentados en [Dea16, DMFS16] que, basándose en técnicas de Revisión de Creencias, atacan los conflictos citados anteriormente (inconsistencia e incoherencia). El primer enfoque (basado en las ideas de Hansson [Han94, Han01]) consiste atacar los conflictos, identificando conflictos mínimos conocidos como *kernels* y removiendo de ellos fórmulas para lograr resolverlos. En el segundo enfoque se utiliza una mirada más global, considerando la relación existente entre los diferentes kernels mediante la agrupación de los mismos en una superestructura denominada *cluster*, la cual es obtenida explotando una relación de solapamiento entre conflictos.

2.2. Revisión de Creencias

El principal objetivo de la *Revisión de Creencias* o la *Teoría de Cambio*, es el modelado de la dinámica del conocimiento. Es decir, la conceptualización del modo en que cambia el conocimiento y se adapta a influencias externas, luego de recibir información nueva.

En los 80', Carlos Alchourrón, David Makinson y Peter Gärdenfors convergieron sus trabajos previos hacia la formalización del *modelo AGM* [AGM85]. Este trabajo es fundacional en la teoría de cambio y proveyó el marco de trabajo (framework) formal más general hasta ese momento para estudios de cambio de creencias.

El fundacional *modelo AGM* especificó un *modelo epistémico* de dinámica de conocimiento; es decir, se definió el formalismo en el cual las *creencias* (o el conocimiento) son representadas y en el cual diferentes clases de operaciones de cambio (*operadores*) pueden ser definidos. A partir de este trabajo, diferentes frameworks para la dinámica de conocimiento han sido propuestos, basándose en este "estilo" de presentación. Las operaciones básicas de cambio definidas en AGM, y tomadas en trabajos subsiguientes, sobre una base de conocimiento KB y una sentencia α , son:

- **Expansión:** expandir la KB mediante α resulta en una nueva KB^f que infiere a α .
- **Contracción:** contraer a KB , por α resulta en una nueva KB^f que no infiere a α .

- **Revisión:** revisar a KB por α resulta en una KB^r , en la cual α es *consistentemente* inferida.

El trabajo [AGM85] se basó en la utilización de teoremas de representación (también llamados caracterizaciones axiomáticas), los cuales sirven para brindar una clara conexión entre las construcciones de los operadores y las propiedades esperadas de los mismos. Tales teoremas se basan en mostrar la relación entre dos componentes fundamentales: los postulados y las construcciones de operadores.

Los postulados de racionalidad [AGM85, Gär88] determinan restricciones que los operadores deben satisfacer. Ellos tratan a los operadores como cajas negras, describiendo sus comportamientos con respecto a la entrada en casos básicos, pero no los mecanismos internos usados. Por otro lado las construcciones se refieren a la contraparte de los postulados, es decir, a definir como conseguir el objetivo esperado por parte del operador. Existen modelos equivalentes posteriores al AGM, tales como: *safe contractions* [AM85], *epistemic entrenchment* [GM88], *sphere systems* [Gro88], y *kernel contractions* [Han94].

Supongamos que KB representa el conocimiento actual y α es una nueva pieza de información. Supongamos que una contracción de KB por α es realizada. Siguiendo el modelo AGM, una *partial meet contraction* está basada en una selección entre los subconjuntos maximales de KB que no implican α . Alternativamente, otra posibilidad es realizar una selección entre los subconjuntos minimales de KB que implican α tal como en *safe contraction* [AM85]. Una variante más general del mismo enfoque fue introducido más tarde y es conocido como *kernel contraction* [Han94]. De esta forma, el proceso de resolución de conflictos es definido como la identificación de tales conflictos mínimos y la posterior remoción de alguna fórmula cada uno de ellos. Logrando así, que la KB se altere lo menos posible. Este estrategia es muy natural para un entorno de aplicación de bases de conocimiento finitas. Ha sido mostrado que las *safe contractions* y las *kernel contractions* son equivalentes a las *partial meet contractions*.

En otros trabajos en el área, otras operaciones adicionales fueron desarrolladas.

La operación de **consolidación de creencias** [Hano1] se aplica cuando una KB es inconsistente. La operación, denotada por $KB!$, se encarga de la remoción (*i. e.*, contracción) de suficientes sentencias de KB hasta llevarla a un estado de consistencia.

La operación de **combinación de creencias** se aplica en situaciones donde el conocimiento es provisto por diversas fuentes y es frecuente que aparezcan conflictos al unir la información. Se puede descomponer la combinación en dos suboperaciones: (a) reunir toda información provista en un único cuerpo, generalmente por unión de conjuntos; y (b) resolver los conflictos que aparezcan en el conocimiento unificado (*i. e.*, *consolidar* el mismo). Como trabajo destacado, Baral *et al.* [BKM91] ha establecido *operadores de combinación*

basados en la unión de todos los conjuntos de creencias, y en la selección de subconjuntos maximalmente consistentes mediante el uso de algún orden dado.

Finalmente, la operación de **integración de creencias** se propone unificar de manera consistente conocimiento provisto por diversas fuentes. Difiere del objetivo de la *combinación* en el sentido que, en los operadores de integración, siempre es posible identificar la procedencia de las distintas piezas de información con las que se cuenta; por lo tanto, en la resolución de conflictos es posible considerar no sólo las características intrínsecas a la información sino también otras informaciones que pueden ser de utilidad, como por ejemplo la cantidad de fuentes que soportan una pieza de información particular, o la credibilidad asociada a las mismas. El principal trabajo relacionado es el de Konieczny y Pino-Pérez [KP02], quienes sentaron las bases más importantes respecto de los operadores de integración. Además proponen una construcción posible para operadores de integración basados en el uso de pre-órdenes totales del conocimiento provisto por las diversas fuentes.

3. Líneas de Investigación y Desarrollo

Esta línea de I+D trata sobre la definición de procesos de integración de ontologías Datalog[±]. Para ello, se utilizan formalismos destinados a la resolución de incoherencias e inconsistencias provenientes de las áreas de Revisión de Creencias y Argumentación. Muchos de estos enfoques, basan su operación en la remoción de información como método de resolución de conflictos. Se propone una resolución alternativa de conflictos que surge de considerar varias ontologías en conjunto mediante la aplicación de *debilitamiento*. En lugar de remover completamente la información conflictiva, se modifica la misma de forma que los conflictos queden resueltos pero no haya que descartar (en su totalidad) la información.

Para esto, distintos ejes deben ser investigados, abarcando desde la definición de incoherencias e inconsistencia en el entorno de ontologías Datalog[±] hasta las posibles aplicaciones que podría tener un método automático de integración de estas ontologías.

3.1. Identificación de Incoherencias e Inconsistencias

A pesar de su creciente popularidad, no ha habido mucho estudio acerca de los aspectos de Representación de Conocimiento en Datalog[±]. Muy poco trabajo se ha hecho acerca de inconsistencias en ontologías Datalog[±]. Peor aún es la situación respecto del concepto de incoherencia.

Uno de los ejes de esta línea de I+D es la definición formal del concepto de incoherencia en Datalog[±], to-

mando como partida esfuerzos similares que han sido realizados para otros formalismos de representación de conocimiento, principalmente Description Logics. Adicionalmente, se procederá a identificar las propiedades que identifican a un conjunto de TGDs como incoherente, y las que hacen que una ontología Datalog[±] se vuelva inconsistente. De esta manera se podrán identificar tales casos, lo que será el primer paso para la posterior resolución de tales problemas.

3.2. Resolución por debilitamiento

Una vez definidos e identificados los conjuntos incoherentes de TGDs y aquellos conjuntos de átomos que provocan inconsistencias en la unión de varias ontologías Datalog[±], se debe proceder a la resolución de estos conflictos. En esta línea de I+D se hará mediante la aplicación de técnicas derivadas de Revisión de Creencias inspiradas en Kernel Contraction [Han94], donde el foco estará puesto en realizar debilitamiento, en lugar del enfoque tradicional que utiliza remoción de información.

En Kernel Contraction clásico, la resolución de incoherencia/inconsistencia se realiza tomando los conjuntos conflictivos mínimos y eligiendo de alguna forma que elemento remover de los mismos para solucionar el problema. En el caso de integración de ontologías Datalog[±] esto es la remoción de ciertos átomos y ciertas TGDs de la unión de todas las ontologías que se está integrando. Aquí se propone tomar ventaja de este enfoque modificando para que se utilicen métodos de debilitamiento de reglas.

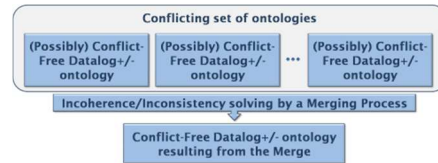
Evidentemente esto no resulta trivial, ya que dado un conjunto conflictivo minimal no es claro cual de sus elementos deberíamos modificar. A su vez, incluso sabiendo cual modificar realizar tal modificación es una tarea compleja, donde surgen preguntas tales como qué modificación realizar. Por ejemplo, es necesario identificar si es mejor realizar debilitamiento mediante el agregado de átomos en el cuerpo de una TGD, o si es preferible modificar los cuantificadores que haya en la misma; o si hay características de las ontologías que hacen que una opción es mejor que otra.

3.3. Integración de ontologías Datalog[±]

En base a los métodos descritos de resolución de incoherencias e inconsistencias se procederá a definir operadores de integración de ontologías Datalog[±] basados en debilitamiento. Además, se darán las propiedades esperadas de tales operadores y se definirán métodos de obtención de operadores de tales características mediante Teoremas de Representación.

3.4. Integración de otras fuentes

Finalmente, se analizará el framework de integración de ontologías Datalog[±] como medio de integración de



otras fuentes de datos. Principalmente nos enfocaremos en la integración automática de Bases de Datos Relacionales.

Para esto, primeramente definiremos métodos para expresar bases de datos relacionales a través de ontologías Datalog[±], tanto los datos en sí como aspectos relacionados al esquema de las mismas, *e. g.*, las dependencias funcionales. Una vez logrado esto se podría utilizar los métodos de integración de ontologías Datalog[±] para obtener una federación de las bases de datos expresadas, ya que la ontología final resultante de la integración brindaría un esquema integrador de las mismas así como los datos que serían parte de la federación, manteniendo a su vez la coherencia de las restricciones de integridad respecto del esquema unificado y la consistencia de los datos almacenados respecto de las dependencias funcionales.

4. Resultados y Objetivos

El objetivo final de esta línea de investigación es el diseño y construcción de la infraestructura necesaria para la realización de procesos de argumentación en una escala masiva valiéndose de datos almacenados en Bases de Datos heterogéneas integradas en Federaciones.

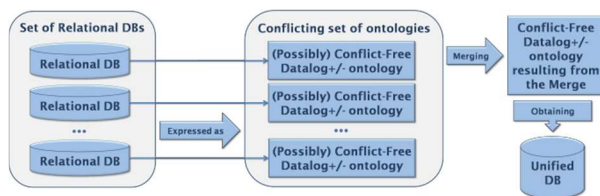
En particular, en el presente trabajo se propone la representación e integración de tales bases de datos a través de la utilización de ontologías Datalog[±], las cuales poseen una expresividad y tratabilidad adecuadas para tal fin. De esta forma, todos los desarrollos que se logren en Datalog[±] serían transferibles a bases de datos (y otras tecnologías estrictamente menos expresivas que Datalog[±] como ciertas Description Logics, *e. g.*, DL-Lite o *EL*).

Un aspecto sumamente importante para la consecución del objetivo final es la definición de métodos automáticos que permitan la creación de vistas unificadas de varias bases de datos como forma de construir las federaciones de bases de datos, de forma que los procesos de argumentación accedan a las mismas para dar soporte a los argumentos. En este trabajo proponemos la consecución de este mediante el uso de integración de ontologías a través de técnicas de debilitamiento.

- Definición de métodos de consolidación de ontologías Datalog[±] que permitan el manejo de inconsistencia e incoherencia. En particular, se propone utilizar un esquema de debilitamiento de reglas y átomos mediante la modificación de los mismos. Para esto, se aprovechará tanto enfoques clásicos

de Revisión de Creencias como Kernel Contraction [Han94, Han01] como así también refinamientos desarrollados en la presente línea de investigación (Cluster Contraction [Dea16, DMFS16]).

- En base a los mecanismos desarrollados procederemos a definir métodos que permitan la integración de ontologías Datalog[±], mediante la consolidación de la unión de ontologías Datalog[±]. Una vez logrado esto, se podría utilizar tales métodos para obtener una federación de las bases de datos expresadas, ya que la ontología final resultante de la integración brindaría un esquema integrador de las mismas, así como los datos que serían parte de la federación, manteniendo a su vez la coherencia de las restricciones de integridad respecto del esquema unificado y la consistencia de los datos almacenados respecto de las dependencias funcionales, como se muestra a continuación.



5. Formación de RR.HH.

En la presente línea de I+D se enmarca el desarrollo de una tesis doctoral para el Doctorado en Ciencias de la Computación de la Universidad Nacional del Sur.

Referencias

- [AGM85] Carlos Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.
- [AK05] Leila Amgoud and Souhila Kaci. An argumentation framework for merging conflicting knowledge bases: The prioritized case. In *ECSQARU*, pages 527–538, 2005.
- [AM85] Carlos Alchourrón and David Makinson. On the Logic of Theory Change: Safe Contraction. *Studia Logica*, 44:405–422, 1985.
- [BKM91] Chitta Baral, Sarit Kraus, and Jack Minker. Combining multiple knowledge bases. *IEEE Trans. Knowl. Data Eng.*, 3(2):208–220, 1991.
- [BQL07] David A. Bell, Guilin Qi, and Weiru Liu. Approaches to inconsistency handling in description-logic based ontologies. In *OTM Workshops (2)*, pages 1303–1311, 2007.
- [CGL12] Andrea Calì, Georg Gottlob, and Thomas Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:57–83, 2012.
- [Dea16] Cristhian Ariel David Deagustini. *Consolidación de ontologías Datalog+*. PhD thesis, Departamento de Ciencias e Ingeniería de la Computación. Universidad Nacional del Sur., 2016.
- [DMFS16] Cristhian Ariel David Deagustini, Maria Vanina Martinez, Marcelo A. Falappa, and Guillermo Ricardo Simari. Datalog+-Ontology Consolidation. *J. Artif. Intell. Res. (JAIR)*, 56:613–656, 2016.
- [FHP⁺06] Giorgos Flouris, Zhisheng Huang, Jeff Z. Pan, Dimitris Plexousakis, and Holger Wache. Inconsistencies, negations and changes in ontologies. In *AAAI*, pages 1295–1300. AAAI Press, 2006.
- [Gär88] Peter Gärdenfors. *Knowledge in Flux: Modeling the dynamics of epistemic states*. MIT Press, 1988.
- [Gär03] Peter Gärdenfors. *Belief revision*, volume 29. Cambridge University Press, 2003.
- [GM88] Peter Gärdenfors and David Makinson. Revisions of knowledge systems using epistemic entrenchment. In *TARK*, pages 83–95, 1988.
- [Gro88] Adam Grove. Two modellings for theory change. *Journal of philosophical logic*, 17(2):157–170, 1988.
- [Han94] Sven Ove Hansson. Kernel contraction. *Journal of Symbolic Logic*, 59(3):845–859, 1994.
- [Han01] Sven Ove Hansson. *A Textbook of Belief Dynamics: Solutions to Exercises*. Kluwer Academic Publishers, Norwell, MA, USA, 2001.
- [KP02] Sébastien Konieczny and Ramón Pino Pérez. Merging information under constraints: A logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.
- [QH07] Guilin Qi and Anthony Hunter. Measuring incoherence in description logic-based ontologies. In *ISWC/ASWC*, pages 381–394, 2007.