Integración de Ontologías Datalog+/-

Cristhian A. D. Deagustini María Vanina Martínez Marcelo A. Falappa
Guillermo R. Simari

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingenierías de la Computación
Universidad Nacional del Sur
Alem 1253 - Bahía Blanca - Buenos Aires - Argentina
(0291) 459-5135
cadd@cs.uns.edu.ar, vanina.martinez@cs.ox.ac.uk, mfalappa@cs.uns.edu.ar, grs@cs.uns.edu.ar

Resumen

En los últimos tiempos, la colaboración y el intercambio de información se han vuelto aspectos cruciales de muchos sistemas. En estos entornos es de vital importancia definir métodos automáticos para resolver conflictos entre el conocimiento compartido por distintos sistemas. Este conocimiento es frecuentemente expresado a través de ontologías que pueden ser compartidas por los sistemas que utilizan el mismo.

En la presente investigación se busca la definición de métodos automáticos de integración de ontologías Datalog+/-. En base a lo logrado en este aspecto se buscará la adaptación del framework desarrollado para su aplicación tanto en la creación de federaciones de Bases de Datos (Data Federation) como en el intercambio de datos (Data Exchange). En estos campos de aplicación estos métodos podrán contribuir brindando la posibilidad de obtener de forma automática un esquema universal que respete tanto como sea posible a los originales manteniendo la coherencia del mismo con respecto a las restricciones de integridad impuestas a los datos, y definiendo que datos pueden ser mantenidos en la federación resolviendo incoherencias en el proceso.

Adicionalmente, se analizarán posibles extensiones a Datalog+/- basadas en formalismos de Argumentación Rebatible, teniendo en cuenta aspectos como la definición de relaciones de inferencia para estas ontologías aumentadas que tengan en cuenta los aspectos no-monótonos de la Argumentación Rebatible, o el impacto de tales relaciones en las conclusiones finales obtenidas y la complejidad de la obtención de las mismas.

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

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 datos federadas**". Director: Guillermo R. Simari. PGI 24/N030. 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.
- "**Desarrollo de Sistemas de Argumentación Masiva sobre Base de Datos Federadas**". Director: Guillermo Simari. Co-director: Cristian Pacífico. PID-UNER 7041. Unidad coordinadora: Universidad Nacional de Entre Ríos.

Este último proyecto se enmarca dentro del ámbito de colaboración entre el Laboratorio de Investigación y Desarrollo (LIDIA) del Dep. de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur; y el Laboratorio de Investigación y Desarrollo en Inteligencia Artificial Concordia (LIDIA Concordia) de la Facultad de Ciencias de la Administración, Universidad Nacional de Entre Ríos.

2. Introducción

En los últimos tiempos la integración e interacción entre diferentes sistemas se ha vuelto muy común, especialmente en entornos colaborativos introducidos desde al arribo de la Web Semántica [BLHL01], *e. g.,* e-commerce.

Sin embargo, la colaboración entre sistemas no siempre puede realizarse de manera directa. Hay ocasiones donde problemas de inconsistencia (e incoherencia) aparecen cuando tomamos el conocimiento provisto por diferentes entidades como uno solo. En formalismos de representación de conocimiento como las ontologías estas inconsistencias suelen aparecer como violaciones a las

restricciones de integridad impuestas a los datos en las fuentes originales del conocimiento, que son violadas en el conocimiento integrado.

La resolución de incoherencias e inconsistencias en conocimiento es admitido como un problema importante que debe ser atacado [GCS10, HvHH+05, HvHtT05, BQL07], especialmente en procesos de integración de conocimiento proveniente de fuentes diversas [BHP09, BKM91, AK05]. En particular; el problema de incoherencia, si bien no ha sido estudiado en profundidad en Datalog+/-, si es conocido en otros formalismos de representación de conocimiento, especialmente en la comunidad de Lógicas Descriptivas (Description Logics - DL), donde ha sido analizado desde diferentes enfoques a través de los años [FHP+06, BB97, KPSH05, QH07].

La obtención de conocimiento coherente y consistente basado en distintas fuentes es un problema que aparece en muchos campos de la IA, *e. g.,* sistemas multi-agentes, y fue atacado desde diferentes enfoques. En particular, mucho trabajo al respecto fue hecho en el campo de Bases de Datos, principalmente en integración de esquemas, donde el objetivo es integrar muchas bases de datos en una sola, resolviendo inconsistencias de forma tal que la calidad de la información sea comprometida en la menor medida posible.

Otro área que ha hecho muchos avances en la integración de fuentes inconsistentes de información es la Revisión de Creencias a través de la definición de procesos de Integración a través de Integración de Creencias(Belief Merging [BHA+01]). A través de los años, diferentes enfoques para la integración se han desarollado, *e. g.,* [BKM91, Cho98, LS98, FKIRS12].

En un proceso de integración se toman varias Bases de Conocimiento (Knowledge Bases - *KBs*) que pueden ser consistentes en sí mismas pero generan conflicto cuando son consideradas juntas, y se obtiene una nueva *KB* que está libre de conflictos y refleja la información de las bases originales tanto como sea posible.

Uno de los trabajos mas influyentes en Belief Merging es el de Konieczny and Pino-Pérez [KP02]. En él, los autores proponen ciertas propiedades lógicas que los operadores de Merging deberían satisfacer, estableciendo a su vez formas de conseguir definir operadores que cumplan estas propiedades, a través de la definición de Teoremas de Representación. Ademas de [KP02], muchos otros enfoques has sido desarrollados para la integración y revisión de bases de datos proposicionales(*e. g.,* [KP02, KM92, LS98, KP11, BMVW10, DJ12]), los cuales brindaron las bases para desarrollos que atacaron estos problemas para (fragmentos de) lógicas de primer orden, principalmente en la familia de lenguajes de Lógicas Descriptivas (Description Logics - DLs) [QLB06, BHP09, MLB05] y Programación Lógica (Logic Programming) [HPW09, DSTW09].

En esta línea de investigación nos enfocamos en la definición de procesos de integración como los presentados que se adecuen a su uso en formalismos de representación de conocimientos aptos para el uso en entornos colaborativos como la Web Semántica. En los últimos años, las bases de conocimiento en forma de ontologias se han vuelto muy populares en estos entornos, ya que proveen formas de representar tanto los datos disponibles en sí como las restricciones impuestos a éstos. Además, el poder expresivo de las ontologías permite realizar tareas importantes en la integración de fuentes de conocimientos [Len02], y juega un rol preponderante en la Web Semántica[BLHL01].

Por lo tanto, es importante definir métodos que permitan la integración de ontologías resolviendo todos los problemas de incoherencia/inconsistencia que puedan aparecer. Esto podría ser el primer paso para la integración de otras fuentes de conocimiento que pueden ser expresadas mediante estas ontologías (*e. g.,* bases de datos relacionales).

En esta línea de investigación nos enfocamos en la integración de ontologías desarrolladas en un lenguaje de ontologías en particular denominado Datalog+/- [CGL12]. La familia de lenguajes de ontologías Datalog permite un estilo modular de representación de conocimiento mediante el uso de reglas de forma similar a la usada en Programación Lógica, y su decidibilidad le permite manejar los volúmenes masivos de datos que podemos encontrar en aplicaciones hoy en día, haciéndola útil en diferentes campos como la consulta de ontologías, extracción de datos en web o intercambio de datos [LMS12]. 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 nos permiten obtener nuevos átomos mediante la activación de las mismas como ser: alumno(X) $\rightarrow$ persona(X), (c) Equality-generating Dependencies: reglas que restringen la generación de átomos, por ejemplo: doctor(D, P) $\wedge$ doctor(D$'$, P) $\rightarrow$ D = D$'$; y (d) Negative Constraints - NCs: reglas que expresan relaciones que no pueden existir entre átomos, *e. g.,*: alto(X) $\wedge$ bajo(X) $\rightarrow \perp$.

Para la deficinión de los procesos de integración proponemos el desarrollo de un framework general para la integración de ontologías Datalog+/- basado en el uso de Kernel Contraction y los operadores de revisión asociados [Han94, Han97, Han01] surgidos del campo de la Revisión de Creencias. Utilizando estos formalismos se podrá solucionar los problemas de incoherencias a inconsistencias que tienden a aparecer en la integración de información proveniente de distintas fuentes.

3. Líneas de Investigación y Desarrollo

Esta línea de investigación se enfoca en la definición de procesos de integración de ontologías Datalog+/- a través del uso de formalismos enfocados en la resolución de incoherencias e inconsistencias provenientes de las áreas de Revisión de Creencias y Argumentación. Para ello distintos ejes deben ser investigados, que van desde la definición de incoherencias e inconsistencia en

el entorno de ontologías Datalog+/- hasta las posibles aplicaciones que un método automático de integración de estas ontologías podría tener.

3.1. Definición de métodos de identificación de Incoherencias e Inconsistencias en Datalog+/-

Datalog+/- se ha vuelto un lenguaje muy popular en los últimos años, y numerosos estudios se han realizado acerca de sus propiedades de decibilidad y la complejidad asociada a la respuesta de consultas en estas ontologías. Sin embargo, 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 (*i. e.,* la imposibilidad de satisfacer cierto conjunto de TGDs sin violar una NC), que ha sido soslayado al punto tal de no haber actualmente una definición formal del mismo.

Uno de los ejes de esta línea de investigación es la definición formal del concepto de incoherencia en Datalog+/-, tomando 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 llevan a que un conjunto de TGDs sea 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 de Incoherencias e Inconsistencias en Datalog+/-

Una vez que se tiene 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 investigación esto se hará mediante la aplicación de técnicas derivadas de la Revisión de Creencias denominadas Kernel Contraction.

Este tipo de técnicas resuelve conflictos de incoherencia/inconsistencia 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. Adicionalmente, se puede pensar en la definición de métodos de debilitamiento de reglas, en lugar de la remoción de las mismas. Esto no es una tarea trivial, ya que hay muchos aspectos a definir, por ejemplo como elegir el mejor candidato entre los átomos o TGDs que pueden eliminarse, lo que a su vez lleva a definir formas (automáticas) de obtener órdenes entre los candidatos. Para esto se procederá a definir operadores de integración de ontologías Datalog+/-, así como se darán las propiedades esperadas de tales operadores y se definirán métodos de obtención de operadores de tales características mediante Teoremas de Representación.

3.3. Posibles extensiones a Datalog+/- mediante Argumentación Rebatible

Otro aspecto a considerar dentro de la línea de investigación es como se beneficiaría Datalog+/- de otros formalismos de representación de conocimientos con propiedades diferentes a aquellas presentes en Datalog+/-. Particularmente, proponemos analizar posibles extensiones a las ontologías Datalog+/- basadas en el uso de formalismos de Argumentación Rebatible como Programación Lógica Rebatible (Defeasible Logic Programming - DeLP) [GS04].

Para tales extensiones se deberán analizar diferentes aspectos. Por ejemplo, la relación de inferencia en estas ontologías aumentadas deberá tener en cuenta los aspectos no-monótonos de la Argumentación Rebatible, llevando a que se modifique la forma en que una consulta es respondida respecto del proceso actual en Datalog+/- , ya que se deberá tener en cuenta el análisis dialéctico llevado a cabo por DeLP antes de responder la misma.

Estas modificaciones en la forma en que información es inferida traerá aparejado un impacto en las inferencias finales de las ontologías en aquellos casos en donde las ontologías no son coherentes o consistentes, proveyendo otra forma de integrar las mismas (*i. e.,* simplemente considerarlas juntas sin importar problemas de incoherencia e inconsistencia,y dejar que el proceso argumentativo los resuelva).

Otro aspecto importante a analizar será el impacto de estos cambios en la decidibilidad y (principalmente) la complejidad del proceso de resolución de consultas en (la extensión de) Datalog+/-.

3.4. Integración de otras fuentes de datos a través del Merging de Ontologías Datalog+/-

Finalmente, se analizará nuestro 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 general de este trabajo de investigación es el diseño y construcción de la infraestructura necesaria para la realización de métodos de integración de ontologías Datalog+/-, así como la identificación de posibilidades de transferencia del framework a la integración de otras fuentes de información como las bases de datos relacionales. Contar con estos métodos permitirá el uso de ontologías (u otras fuentes) de manera segura en el desarrollo de sistemas colaborativos. Por ejemplo, se podrían definir nuevas arquitecturas para Sistemas de Soporte de Decisiones (DSS) que acceda a información proveniente de diferentes entidades de manera transparente sin tener problemas de inconsistencias o incoherencias. Una extensión análoga podría definirse para los Sistemas de Recomendación (RS), en especial aquellos que basan su funcionamiento sobre una base de conocimiento y que utilizan tecnologías de Web Semántica, es decir, Sistemas de Recomendación Semánticos.

Respecto de los resultados y objetivos alcanzados, se ha avanzado en la definición e identificación de los casos de incoherencia e inconsistencia en Datalog+/-. A su vez, se ha comenzado la definición de operadores de integración de ontologías Datalog+/- basados en Kernel Contraction, estableciendo las propiedades de los mismos y como obtenerlos. Estos resultados se presentarán en congresos nacionales e internacionales.

Adicionalmente, los operadores definidos están siendo implementados actualmente, para así poder probar su eficacia mediante pruebas en la integración de ontologías ya desarrolladas adaptadas al estilo de representación de conocimiento de Datalog+/-. También se esta avanzando en la definición de procesos alternativos de integración de ontologías que utilizan formalismos no-monótonos como la Argumentación Rebatible para la resolución de conflictos, habiendo establecido ya como los mismos podrían usarse sobre un conjunto de ontologías Datalog+/-. En lo subsecuente se procederá a analizar las propiedades de un framework de estas características, así como las posibilidades computacionales del mismo. Dentro de las propiedades a analizar del mismo se plantea como objetivo futuro realizar un estudio comparativo de nuestra propuesta respecto de un formalismo proveniente del área de Bases de Datos cuyo uso para la resolución de inconsistencias ha sido muy extendido: Consistent Query Answering(CQA) [ABC99]. En particular, se analizará la completitud y la sensatez de nuestra propuesta respecto de CQA, así como también las posibles extensiones en las inferencias que nuestro framework podría tener respecto del mismo.

5. Formación de Recursos Humanos

En la presente línea de investigación se enmarca el desarrollo de una tesis de posgrado en el Doctorado en Ciencias de la Computación del Departamento de Ciencias e Ingenierías de la Computación de la Universidad Nacional del Sur.

Referencias

[ABC99] Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki. Consistent query answers in inconsistent databases. In Victor Vianu and Christos H. Papadimitriou, editors, *PODS*, pages 68–79. ACM Press, 1999.

[AK05] Leila Amgoud and Souhila Kaci. An argumentation framework for merging conflicting knowledge bases: The prioritized case. In *Proc. of ECSQARU 2005*, pages 527–538, 2005.

[BB97] Domenico Beneventano and Sonia Bergamaschi. Incoherence and subsumption for recursive views and queries in object-oriented data models. *Data Knowl. Eng.*, 21(3):217–252, 1997.

[BHA+01] Isabelle Bloch, Anthony Hunter, Alain Appriou, André Ayoun, Salem Benferhat, Philippe Besnard, Laurence Cholvy, Roger M. Cooke, Frédéric Cuppens, Didier Dubois, Hélène Fargier, Michel Grabisch, Rudolf Kruse, Jérôme Lang, Serafín Moral, Henri Prade, Alessandro Saffiotti, Philippe Smets, and Claudio Sossai. Fusion: General concepts and characteristics. *Int. J. Intell. Syst.*, 16(10):1107–1134, 2001.

[BHP09] Elizabeth Black, Anthony Hunter, and Jeff Z. Pan. An argument-based approach to using multiple ontologies. In *SUM*, pages 68–79, 2009.

[BKM91] Chitta Baral, Sarit Kraus, and Jack Minker. Combining multiple knowledge bases. *IEEE Trans. Knowl. Data Eng.*, 3(2):208–220, 1991.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):3443, 2001.

[BMVW10] Richard Booth, Thomas Andreas Meyer, Ivan José Varzinczak, and Renata Wassermann. Horn belief change: A contraction core. In *ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 1065–1066, 2010.

[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. *J. Web Sem.*, 14:57–83, 2012.

[Cho98] Laurence Cholvy. Reasoning about merged information. In Didier Dubois, Henri Prade, Dov M. Gabbay, and Philippe Smets, editors, *Belief Change*, volume 3 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 233–263. Springer Netherlands, 1998.

[DJ12] James P. Delgrande and Yi Jin. Parallel belief revision: Revising by sets of formulas. *Artif. Intell.*, 176(1):2223–2245, 2012.

[DSTW09] James P. Delgrande, Torsten Schaub, Hans Tompits, and Stefan Woltran. Merging logic programs under answer set semantics. In Patricia M. Hill and David Scott Warren, editors, *ICLP*, volume 5649 of *Lecture Notes in Computer Science*, pages 160–174. Springer, 2009.

[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.

[FKIRS12] Marcelo Alejandro Falappa, Gabriele Kern-Isberner, Maurício Reis, and Guillermo Ricardo Simari. Prioritized and non-prioritized multiple change on belief bases. *Journal of Philosophical Logic*, 41(1):77–113, 2012.

[GCS10] Sergio Alejandro Gómez, Carlos Iván Chesñevar, and Guillermo Ricardo Simari. Reasoning with inconsistent ontologies through argumentation. *Applied Artificial Intelligence*, 24(1&2):102–148, 2010.

[GS04] Alejandro Javier García and Guillermo Ricardo Simari. Defeasible logic programming: An argumentative approach. *TPLP*, 4(1-2):95–138, 2004.

[Han94] Sven Ove Hansson. Kernel contraction. *J. Symb. Log.*, 59(3):845–859, 1994.

[Han97] Sven Ove Hansson. Semi-revision (invited paper). *Journal of Applied Non-Classical Logics*, 7(2), 1997.

[Han01] Sven Ove Hansson. *A Textbook of Belief Dynamics: Solutions to Exercises.* Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[HPW09] Julien Hué, Odile Papini, and Eric Würbel. Merging belief bases represented by logic programs. In Claudio Sossai and Gaetano Chemello, editors, *ECSQARU*, volume 5590 of *Lecture Notes in Computer Science*, pages 371–382. Springer, 2009.

[HvHH+05] Peter Haase, Frank van Harmelen, Zhisheng Huang, Heiner Stuckenschmidt, and York Sure. A framework for handling inconsistency in changing ontologies. In *Proc. of ISWC 2005*, pages 353–367, 2005.

[HvHtT05] Zhisheng Huang, Frank van Harmelen, and Annette ten Teije. Reasoning with inconsistent ontologies. In *Proc. of IJCAI 2005*, pages 454–459, 2005.

[KM92] Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. *Artif. Intell.*, 52(3):263–294, 1992.

[KP02] Sébastien Konieczny and Ramón Pino Pérez. Merging information under constraints: A logical framework. *J. Log. Comput.*, 12(5):773–808, 2002.

[KP11] Sébastien Konieczny and Ramón Pino Pérez. Logic based merging. *J. Philosophical Logic*, 40(2):239–270, 2011.

[KPSH05] Aditya Kalyanpur, Bijan Parsia, Evren Sirin, and James A. Hendler. Debugging unsatisfiable classes in owl ontologies. *J. Web Sem.*, 3(4):268–293, 2005.

[Len02] Maurizio Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246, 2002.

[LMS12] Thomas Lukasiewicz, Maria Vanina Martinez, and Gerardo I. Simari. Inconsistency handling in datalog+/- ontologies. In *Proc. of ECAI*, pages 558–563, 2012.

[LS98] Paolo Liberatore and Marco Schaerf. Arbitration (or how to merge knowledge bases). *IEEE Trans. Knowl. Data Eng.*, 10(1):76–90, 1998.

[MLB05] Thomas Meyer, Kevin Lee, and Richard Booth. Knowledge integration for description logics. In *In Proceedings of the 7 th International Symposium on Logical Formalizations of Commonsense Reasoning*, pages 645–650. AAAI Press, 2005.

[QH07] Guilin Qi and Anthony Hunter. Measuring incoherence in description logic-based ontologies. In *ISWC/ASWC*, pages 381–394, 2007.

[QLB06] Guilin Qi, Weiru Liu, and David A. Bell. Knowledge base revision in description logics. In Michael Fisher, Wiebe van der Hoek, Boris Konev, and Alexei Lisitsa, editors, *JELIA*, volume 4160 of *Lecture Notes in Computer Science*, pages 386–398. Springer, 2006.