Completitud de los Métodos de Acceso a Datos Basado en Ontologías: Enfoques, Propiedades y Herramientas

Sergio Alejandro Gómez, Pablo Rubén Fillottrani
Laboratorio de I&D en Ingeniería de Software y Sistemas de Información (LISSI)
Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
San Andrés 800 - Campus de Palihue, (8000) Bahía Blanca, {sag,prf}@cs.uns.edu.ar
Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

RESUMEN

Exploramos el acceso a datos basado en ontologías (OBDA). Una ontología brinda una visión conceptual de un repositorio de datos relacional. La instancia relacional está representada con una base de datos relacional mientras que el esquema de la base está representado en el lenguaje OWL2. OWL2 tiene tres *perfiles* que hacen ciertas concesiones en la capacidad expresiva para salvaguardar la eficiencia computacional de las operaciones pero comprometiendo la completitud del razonamiento. Estudiaremos cómo es posible extender tales capacidades de representación para mejorar la completitud de los métodos de OBDA.

Palabras clave: Acceso a datos basado en ontologías, Description Logics, Representación de conocimiento.

CONTEXTO

Esta línea de investigación se ejecuta en el contexto de un Proyecto de Grupo de Investigación de la Universidad Nacional del Sur y de un Proyecto de Innovación y Transferencia en Áreas Prioritarias de la Provincia de Buenos Aires (PIT-AP-BA) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA) llamado "*Herramientas para el desarrollo y la entrega de servicio públicos digitales de acción social para municipios bonaerenses*".

1. INTRODUCCIÓN

Una ontología es una formalización de unaparte de un dominio de aplicación dando una visión conceptual de los repositorios de datos y que se ha venido haciendo más y más popular desde el 2001 aproximadamente, en particular en las áreas de la Integración de Aplicaciones Empresariales, la Integración de Datos y la Web Semántica [Calvanese et al., 2006].

El acceso a datos basado en ontologías (OBDA) [Cali et al., 2012b; Calvanese et al., 2013] es visto como un ingrediente clave en la nueva generación de sistemas de información. En el paradigma de OBDA, una ontología define un esquema global de alto nivel de fuentes de datos (existentes) y brinda un vocabulario para consultas de usuario. Un sistema de OBDA reescribe tales consultas y ontologías en el vocabulario de las fuentes de datos y luego delega la evaluación real de la consulta a un sistema adecuado como por ejemplo una base de datos relacional o un motor Datalog [Kontchakov et al., 2013].

Como plantea [Stoilos, 2014], si bien la utilización de ontologías OWL provee un marco para la conceptualización formal y semántica de las fuentes de datos subyaciendo muchas aplicaciones modernas, el poder expresivo de OWL DL tiene un alto precio respecto a su complejidad computacional, aún luego del diseño de implementaciones con modernas optimizaciones, los razonadores OWL DL

todavía no fueron capaces de lidiar con grandes bases de datos conteniendo miles de millones de registros. Como una consecuencia de ello, en aplicaciones del mundo real, los desarrolladores a menudo emplean sistemas de respuestas de consultas escalables y eficientes que soportan solamente un perfil de OWL2 en alguno de sus perfiles, sea OWL2-EL, OWL2-QL uOWL-RL. En consecuencia, ya que OWL2 es compatible hacia atrás con OWL, tales sistemas cargan una ontología OWL pero ignoran todas sus partes que caen fuera del fragmento que soportan. En consecuencia, son *incompletas*; es decir, para algunas ontologías, consultas e instancia debase de datos fallarán en computar todas las respuestas ciertas. A pesar de que la escalabilidad es muy atractiva, la resolución de consultas incompleta no es aceptable en ciertas aplicaciones críticas como por ejemplo los dominios de la salud y la defensa militar; por lo tanto, el mejoramiento de la completitud por medio de la computación de tantas respuestas *perdidas* como sea posible sin afectar la performance sería muy beneficioso para muchas aplicaciones.

2. LÍNEAS DE INVESTIGACIÓN

Lenguajes de representación de ontologías: El estándar moderno de representación de conocimiento en la Web Semántica está dado por el lenguaje OWL2 [Hitzler et al., 2012]. OWL2 utiliza una sintaxis XML para el intercambio de datos pero su semántica está basada en las Lógicas para la Descripción (DL) [Baader et al., 2003]. Una ontología DL consiste de dos conjuntos finitos y mutuamente disjuntos: una Tbox que introduce la terminología y una Abox que contiene aserciones acerca de objetos particulares en el dominio de aplicación. En particular, OWL2 DL tiene tres *perfiles* con complejidad tratable: OWL2-EL, OWL2-QL y OWL2-RL dependiendo de la lógica para la descripción subyacente que le da significado [Hitzler et al., 2012]. Respecto de su complejidad temporal, OWL2-EL tiene tiempo polinomial para esquema y datos, OWL2-QL permite respuestas a consultas en forma rápida (LOGSPACE) usando sistemas gestores de base de datos relacional vía SQL y OWL2-RL permite respuestas a consultas en tiempo polinomial usando bases de datos extendidas con reglas. En base a esto, cada perfil es útil para un cierto tipo de problema: OWL2-EL es útil para ontologías con una parte conceptual grande, OWL2-QL es útil para grandes datasets almacenados en RDBs, y OWL2-RL es útil para grandes conjuntos de datos almacenados como triplas RDF.

Acceso a datos basado en ontologías: En el acceso a datos basado en ontologías (OBDA) una ontología define un esquema global de alto nivel de una base de datos brindando un vocabulario para consultas, reescribiendo las consultas y ontologías en el vocabulario de las fuentes de datos y luego delegando la evaluación a un motor relacional o Datalog [Kontchakov et al., 2013]. Los sistemas de OBDA son importantes porque (i) brindan una vista conceptual del alto nivel de los datos, (ii) brindan al usuario un vocabulario conveniente para consultas, (iii) permiten al sistema enriquecer datos incompletos con conocimiento del dominio, y (iv) soportan consultas sobre múltiples fuentes de datos heterogéneas. Se pueden distinguir varios tipos de OBDA dependiendo del poder expresivo de las DL involucradas: (a) OBDA con bases de datos (por ejemplo, las lógicas de la familia de DL-Lite [Calvanese et al., 2013] y su implementación en XML OWL2-QL permiten la reducción de consultas conjuntivas sobre ontologías a consultas de primer orden sobre bases de datos relacionales estándar); (b) OBDA sobre motores Datalog (por ejemplo, las DL de la familia EL [Lutz et al., 2009] y su implementación XML OWL2 EL, Horn-*SHIQ* y Horn-*SROIQ* soportan un reducción a Datalog, y (c) OBDA con DLs expresivas (como *ALC* o *SHIQ* que requieren técnicas especiales para analizar consultas conjuntivas).

OBDA con bases de datos relacionales: Las lógicas de la familia DL-Lite [Calvanese et al., 2013] (y por ende el perfil OWL2 QL [Motik et al., 2012]) permiten una reducción de las consultas conjuntivas sobre ontologías a consultas de primer orden sobre bases de datos relacionales estándar. Así, una de las nociones fundamentales en OBDA es la de *reescritura de consultas*; en ella, dada una ontología (*T*,*A*), el usuario formula una consulta $q(x)$ expresada en el vocabulario una terminología *T* de una ontología, la tarea del sistema de OBDA es la de reescribir $q(x)$ y *T* en una nueva consulta equivalente $q'(x)$ expresada en el vocabulario de los datos *A* tal que para cualquier conjunto de datos, las respuestas de $q(x)$ sobre la *(T,A)* son las mismas que las respuestas de $q'(x)$ sobre *A*. Así, el problema de consultar los datos *A* (cuya estructura es desconocida al usuario) en términos de una ontología *T* (accesible por el usuario) se reduce al problema de consultar *A* directamente; así cuando la *A* es modelada con la instancia relacional de una base de datos relacional, tal tarea se puede realizar muy eficientemente aprovechando los mecanismos de optimización de consultas mediante métodos estándar que permiten reexpresar un consulta de primerorden $q'(x)$ como un conjunto de consultas SQL. Otros métodos duales al anterior permiten realizar virtualización de Aboxes; en tal caso, la instancia relacional es reexpresada como una Abox *A* y se utiliza un motor DL tradicional para evaluar la consulta. La relación entre los conceptos de la ontología y los datos de la base de datos relacional se expresan con un mapeo *M* (obtenidos manualmente o en forma semiautomática),que es un conjunto de reglas $S(x) \square \varphi(x,z)$ donde *S* es un nombre de concepto o rol de la ontología y $\varphi(x,z)$ es una conjunción de átomos con relaciones de bases de datos (almacenadas o en vistas) y un filtro relacional. Sin embargo y con el objetivo de mantener la complejidad temporal de los algoritmos involucrados en forma tratable, ciertas restricciones de integridad de algunos modelos conceptuales de datos (e.g. entidad-relación) no pueden ser representadas en OWL2 QL (e.g. disyunciones en el miembro derecho deaxiomas de inclusión -lo que impide modelar herencia múltiple- o restricciones numéricas en la cardinalidad de relaciones o roles). Otros problemas del enfoque de reescritura de consultas se da cuando la reescritura $q'(x)$ es muy grande para ser manejada exitosamente por el gestor relacional de bases de datos donde serán ejecutadas. El mapeo *M* puede contener las reglas que relacionan los términos de la ontología con el esquema de la base de datos. Entonces, dada una consulta $q(x)$ se puede obtener una reescritura $q'(x)$, la cual se puede desdoblar en una consulta SQL usando evaluación parcial. La evaluación parcial aplica resolución SLD a $q'(x)$ y al mapeo *M* y retorna aquellas reglas cuyos cuerpos contienen sólo átomos de la base de datos. Así, cada regla de $q'(x)$ resulta en una consulta SQL de tipo Select-Project-Join que se envía al manejador de bases de datos relacional para su ejecución. Es de notar, que a pesar de que el tema de la reescritura de consultas sobre DL-Lite (y en consecuencia sobre OWL2-QL) está bastante maduro, como DL-Lite no permite modelar todas las restricciones impuestas por ciertos modelos conceptuales (por ejemplo, al no poder usarse la disyunción en los miembros derechos de los axiomas de inclusión de DL-Lite, no es posible modelar *restricciones de cubrimiento* como que dos subclases son completas, i.e. no puedehaber una tercera subclase de la superclase; otras restricciones no modelables implica decir que un atributo es funcional, y también las restricciones de cardinalidad tampoco pueden representarse en esta familia de DL requiriendo un lenguaje más expresivo [Kontchakov et al.]).El problema reside en que agregar tales restricciones destruye la propiedad de *reescritura de primer orden* con lo que al final terminaría impactando en la eficiencia del sistema.

OBDA con bases de datos no-relacionales: El paisaje del tema Bases de Datos se ha diversificado significativamente durante la última década, resultando en el surgimiento de

una variedad de bases de datos no relacionales (NoSQL), por ejemplo bases de datos XML y documentos JSON, almacenes clave-valor y bases de datos en forma de grafo [Botoeva et al., 2016]. [Harris and Seaborne, 2013] están investigando generalizaciones del marco OBDA para permitir la consulta de bases de datos arbitrarias a través de ontologías mediadoras utilizando el sistema de *Ontop* para OBDA y consultas SPARQL sobre la base de documentos *MongoDB* [Botoeva et al., 2016].

Problema de la completitud del razonamiento en el acceso a datos basado en ontologías: Como explicamos previamente, los perfiles de OWL2 para modelar fuentes de datos sacrifican completitud para mantener un nivel de eficiencia aceptable. Sin embargo, en aplicacionescomo la salud o la milicia, no es posible darse el lujo de perder respuestas a consultas. Repasamos ataques al problema encontrados en la literatura.

[Ma et al., 2006] presentan un *benchmark* para evaluar las capacidades de inferencia de sistemas ontológicos para los lenguajes OWL Lite y OWL DL. Cómo proveer aproximaciones escalables y de calidad para la resolución de consultas en DLs expresivas es un importante problema en representación de conocimiento. Debido al peor caso de complejidad para el problema de razonamiento de los perfiles expresivos de OWL2, muchas veces un cambio pequeño realizado por un ingeniero de conocimiento produce unas demoras de complejidad muy acentuadas. De la misma forma, cambiar de un razonador a otro puede producir cambios significativos en los tiempos de clasificación. [Goncalves et. al, 2012] investigan la identificación en una forma sistemática de los llamados *hot spots* (puntos calientes), que son subconjuntos difíciles de las ontologías, que, al ser removidos, producen una mejora significativa en la eficiencia. [Pan and Thomas, 2007] brindan una aproximación que garantiza sensatez y que transforma una ontología OWL expresada en una DL más expresiva en una aproximación maximal en una DLtratable. [Pan et al., 2009] estudian el mismo problema pero tratando de mantener la completitud. [Tserendorj et al., 2008] presentan un acercamiento al razonamiento aproximado para ontologías, basado en el sistema KAON2, llamado SCREECH para razonar con ontologías OWL DL en la forma de su compilación en Datalog disyuntivo. El balance que logran se da en perder completitud en el razonamiento para ganar eficiencia.

3. RESULTADOS ESPERADOS

El *objetivo general* de este Plan de Trabajo involucra investigar mejoras en las capacidades de representación de conocimiento y razonamiento con ontologías en relación a los métodos, algoritmos y herramientas del acceso a datos en bases de datos basados en ontologías.

El conjunto de *objetivos particulares* de este Plan de Trabajo comprende la investigación de las extensiones a lenguajes de representación de conocimiento para modelar ontologías y su relación con el acceso a datos basado en ontologías. Esto integra los objetivos particulares de: (i) estudiar los lenguajes aptos para realizar acceso a datos basados en ontologías, (ii) estudiar las tecnologías asociadas (manejadores de bases de datos tradicionales y no tradicionales, razonadores DL, plataformas de implementación), (iii) estudiar y ampliar los límites del poder de representación y las clases de problemas atacables, (iv) entender las propiedades de los acercamientos planteados para (v) determinar las posibilidades de implementación computacional con miras a (vi) utilizar tal implementación en mejorar el desarrollo de las tareas mencionadas arriba en relación a la ingeniería de conocimiento de ontologías OWL2-EL, OWL2-QL y OWL2-RL en el contexto de la iniciativa de la Web Semántica.

4. FORMACIÓN DE RECURSOS HUMANOS

En relación a este plan de trabajo se está desarrollando la dirección de una tesis de maestría y una tesina de grado. Se espera ampliar el número de becarios y tesistas de grado.

5. BIBLIOGRAFÍA

[Baader et al., 2003] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. PatelSchneider, editors. The Description Logic Handbook Theory, Implementation and Applications. Cambridge University Press, 2003.

[Botoeva et al., 2016] E. Botoeva, D. Calvanese, B. Cogrel, M. Rezk, G. Xiao. OBDA beyond relational DBs: A study for MongoDB. In 29th Int. Workshop on Description Logics, volume 1577 of CEUR Electronic Workshop Proceedings, 2016.

[Cali et al., 2012b] A. Cali, G. Gottlob, T. Lukasiewicz. A general Datalog-based framework for tractable query answering over ontologies. Web Semantics: Sciences, Services and Agents on the WWW, No. 14, pp. 57-83, 2012.

[Calvanese et al., 2013] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data Complexity of Query Answering in Description Logics. Artificial Intelligence, Volume 195, February 2013, pp. 335-360.

[Eiter et al., 2012b] T. Eiter, M. Ortiz, M. Simkus, T. Tran, G. Xiao. Query Rewriting for Horn-SHIQ plus Rules. In AAAI-2012, Toronto, Canada, July 22-26, 2012.

[Goncalves et. al, 2012] R. Goncalves, B. Parsia, U. Sattler. Performance Heterogeneity and Approximate Reasoning in Description Logic Ontologies, International Semantic Web Conference (ISWC 2012), pp. 82-98, 2012.

[Harris and Seaborne, 2013] S. Harris and A. Seaborne. SPARQL 1.1 Query Language: W3C Recommendation 21 March 2013,

[Hitzler et al., 2012] P. Hitzler, M. Krötzsch, B. Parsia, P. Patel-Schneider, and S. Rudolph. OWL 2 Web Ontology Language Primer (Second Edition), W3C Recommendation 11 December 2012.

[Kontchakov et al., 2013] R. Kontchakov, M. Rodriguez-Muro, M. Zakharyaschev. Ontology-Based Data Access with Databases: A Short Course. Reasoning Web: Semantic Technologies for Intelligent Data Access, Vol. 8067, LNCS, pp. 194-229, Springer, 2013.

[Lutz et al., 2009] C. Lutz, D. Toman, and F. Wolter. Conjunctive query answering in the description logic EL using a relational database system, (IJCAI 2009), pp. 2070-2075, 2009.

[Motik et al., 2012] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 Web Ontology Language: Profiles (Second Edition) - W3C Recommendation 11 December 2012.

[Pan et al., 2009] J. Pan, E. Thomas, Y. Zhao. Completeness Guaranteed Approximation for OWL DL Query Answering, Proc. of DL, 477, 2009.

[Stoilos, 2014] G. Stoilos. Ontology-Based Data Access Using Rewriting, OWL 2 RL Systems and Repairing. In V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 317-332, 2014.

[Tserendorj et al., 2008] T.Tserendorj, S.Rudolph, M.Krötzsch, P.Hitzler. Aproximate OWL-Reasoning with SCREECH, Technical Report, Wright State University CORE Scholar, Computer Science and Engineering Faculty Publications, 2008.