Knowledge-Based Systems 35 (2012) 144–159

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Ontology-driven relational query formulation using the semantic and assertional capabilities of OWL-DL

K. Munir *, M. Odeh, R. McClatchey

*Centre for Complex Cooperative Systems, Faculty of Environment and Technology, University of the West of England, BS16 1QY, Bristol, United Kingdom*

ARTICLE INFO

*Article history:*
Received 14 November 2011
Received in revised form 26 March 2012
Accepted 19 April 2012
Available online 3 May 2012

*Keywords:*
Information systems
Domain knowledge
Data retrieval
Ontology
Query formulation

ABSTRACT

This work investigates the extent to which domain knowledge, expressed in a domain ontology, can assist end-users in formulating relational queries that can be executed over a complex relational database. In this regard, an ontology-driven query formulation architectural framework has been devised, namely OntoQF, that implements a two-phased approach – the pre-processing and translation phases. In the pre-processing phase, a new database-to-ontology transformation approach has been synthesised where domain ontology is populated and enriched with problem domain concepts and semantic relationships specified using OWL-DL. Once domain ontology has been formulated, end-users can write sophisticated ontology-based queries that are then translated, in the translation phase, into the corresponding relational query statements. In order to validate the correctness of translating single or multiple OWL-DL constructs into their corresponding relational ones, a set of test cases have been derived from the medical domain. Our results demonstrated that the OntoQF framework enriches domain ontology and its associated algorithms drive the process of relational query formulation without the need to both replicate transactional data into the associated domain ontology and have knowledge of the underlying database schema.



## 1. Introduction

In recent years, the substantial increase in the use of medical knowledge discovery and decision support applications has required clinical researchers to write complex database queries. The end-users of these applications are normally unaware of the semantic relationships between the concepts stored in their respective databases. This problem grows with an increase in the structural complexity of the underlying data. Despite the variety of approaches available so far, three major concerns remain: (1) what type of requests can a specific system handle? (2) how can visual interfaces be provided to generate data requests? and (3) how can users be assisted in formulating queries in order to retrieve the most relevant information? Today, information technology has been widely adopted in resolving the first two problems by providing some solutions using graph theory, with the provision of visual tools to generate specific queries (e.g. in [1,2]). However, little has been achieved in providing users with query formulation services in order to retrieve relevant information.

In order to provide query formulation services, new mechanisms for generating queries are required. In an attempt to assist domain experts to generate domain-specific queries, application domain specific ontologies [3] can be used. In general, an ontology represents a shared, agreed and detailed model (or set of concepts) of a certain problem domain [4]. One major advantage of using domain ontology is its ability to define a semantic model of the data combined with the associated domain knowledge. Ontologies can also be used to define links between different types of semantic knowledge. Hence, ontologies can be used in formulating some data searching strategies. This work is concerned with the use of medical domain knowledge specified in an ontology, as a foundation for assisting clinical researchers in generating relational database queries. Based on this, the following research questions have been answered in this paper:

Q.1: *to what extent can a relational database schema be utilised to enable the process of formulating ontology based queries*? and
Q.2: *how can domain specific knowledge be used to enable the process of query formulation and in what format should such domain knowledge be structured*?

Following the relational database paradigm, a relational database model establishes a link between both end-user and domain requirements with the aim of describing a logical structure of the underlying database. SQL, the relational structured query language that is based on the Relational Algebra (RA), is the standard language used to query relational databases. The medical ontologies used in this

* Corresponding author. Tel.: +44 7809682467.
*E-mail address:* kamran.munir@cern.ch (K. Munir).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.04.020

research are specified in OWL-DL [5], therefore, query formulation methods need to interpret and transform OWL-DL expressions into relational expressions, as shown in Fig. 1b. In this regard, we have also addressed the following research question:

Q.3: *what mappings are required to translate OWL-DL expressions to relational ones*?

The work reported here has been applied to the Health-e-Child (HeC) project [6]. The Health-e-Child (HeC) project aims to develop an integrated platform for European paediatrics, enabling data integration between spatially distributed clinicians and bringing together information produced in multiple hospitals. Its cornerstone is the integration of information across biomedical abstractions, whereby all layers of biomedical information can be 'vertically integrated' [6]. The provision of semantic query formulation services in HeC aims at the provision of semantics-driven query formulation services for the clinical researchers and medical knowledge discovery applications. The following are two major uses of an ontology based query formulation system in the HeC project:

- a user can formulate a query without specific knowledge of the information structure and access mechanisms of the underlying data source; and that
- clinical researchers can focus their data query requests around a particular genre of information, e.g., clinical observations, genetic information or laboratory test data, and therefore obtain a rather narrow and fragmented view of the individual patient that they are examining or the disease that they are investigating.

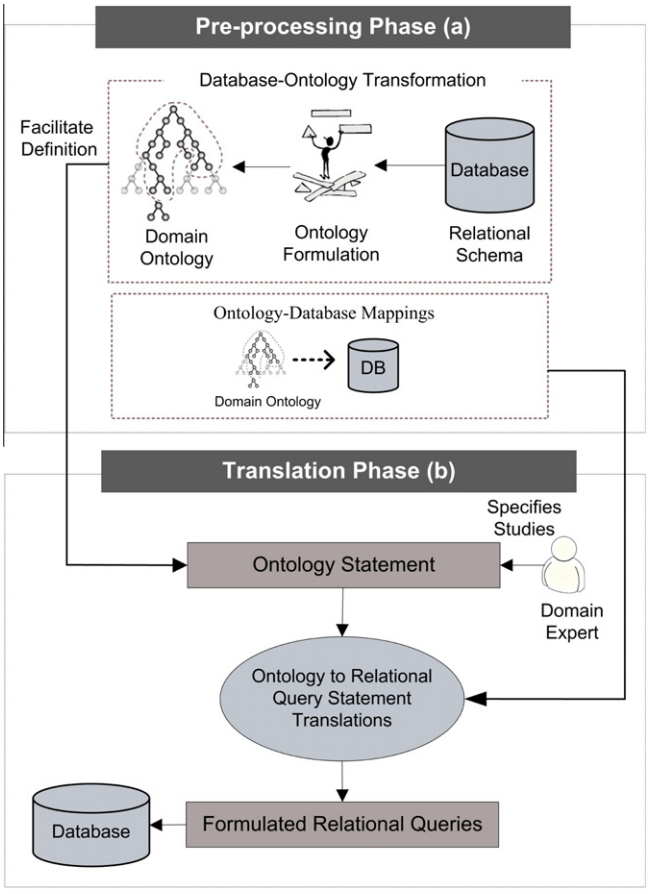


**Fig. 1.** The pre-processing and translation phases of ontology-driven query formulation approach.

For example, a clinical researcher may want to perform a study on patients' *infections*. In doing so, an associated domain ontology-driven system should recognise that *bacterial* and *viral* are types of infections and that *meningitis*, *rat bite fever*, *scarlet fever*, etc. are sub-types of *bacterial* infection. Here, the query formulation system should also identify how the associated data are structured in the underlying database in order to transparently retrieve the resultant dataset. Similarly, for a particular clinical study of *female* patients diagnosed with the medical disorder *double-vision* but who were using *anti-depression* drugs, the associated domain ontology based system should recognise that double vision is a type of clinical test with the possible values *true* (to confirm the affliction) or *false* (to disprove its existence) and also to check for the medical history of female patients prescribed with antidepression drugs to retrieve the desired resultant dataset.

The query formulation methods presented in this paper are developed under two basic assumptions, i.e. enabling the developed query formulation methods to (a) be flexible in terms of accommodating changes in the underlying database schema; and (b) provide access to existing relational database without manipulating or replicating the transactional data; and in particular, where it is not practically feasible to store all data as part of a certain domain ontology especially for systems with large amounts of data.

## 2. Related work

In our perspective, an ontology that describes the logical structure of a domain, its concepts, and the relationships, is used for relational query formulation. The related database to ontology formulation approaches, e.g. presented in D2R-MAP [7], extended D2R [8], R2O [9], VisAVis [10,11] consider the database-ontology mappings while assuming the existence of both a relational database (DB) and a domain ontology. These mapping approaches relate each construct in the relational database to a construct in the ontology and ignore unrelated constructs between the two. The R2O, for example, is an extensible and declarative language to describe mappings between relational DB schemas and ontologies implemented in RDF (S) [12] or OWL [13]. The most important aspect of this approach is the use of the database schema and the ontology as they are, and it defines a declarative specification of the mappings between its modelling components. The R2O is an extension of recent approaches like D2R-MAP and extended D2R. Most of these existing database-ontology formulation approaches do not provide an exact representation of the domain-metadata in an ontology that could be used to generate the respective database relations. While using domain ontology to formulate relational queries, some of the basic rules to generate domain ontology from relational schema remain the same as reported in [7,8]. However, such transformation approaches do not further assist in specifying concept restrictions to generate precise database queries. In our approach, the existing mapping schemes are extended to support query formulation needs by introducing further semantic groupings with respect to cardinality relationships between *Domain Metadata*. Here the term *Domain Metadata* refers to data stored in those database tables which contain *entity* related information. For example, in a M:N relationship between *Customer* and *Product* entities, both *Customer* and *Product* tables store entity related data, and the database table that is generated as a result of creating a M:N relationship between *Customer* and *Product* entities is named as *transaction table* (*that stores transactional data*/*logs*) or *bridge*/*join table*.

In recent years, some ontology based query formulation systems are reported that use visual representations to express related data requests (e.g. TAMBIS [14], GRQL [15], SEWASIE [1], SHOE [16], SEAL [17], in [2], Ontogator [18], OntoViews [19], etc.). In most of these