eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

An Ontology Based Method to Solve Query Identifier Heterogeneity in Post-Genomic Clinical Trials

Alberto ANGUITA[a], Luis MARTÍN[a], José CRESPO[a] and Manolis TSIKNAKIS[b]
[a]*Biomedical Informatics Group, Artificial Intelligence Laboratory, School of Computer Science, Universidad Politécnica de Madrid Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain*
[b]*Center of eHealth Technologies, Institute of Computer Science, Foundation for Research and Technology - Hellas GR-71110 Heraklion, Crete, Greece*

Abstract. The increasing amount of information available for biomedical research has led to issues related to knowledge discovery in large collections of data. Moreover, Information Retrieval techniques must consider heterogeneities present in databases, initially belonging to different domains—e.g. clinical and genetic data. One of the goals, among others, of the ACGT European is to provide seamless and homogeneous access to integrated databases. In this work, we describe an approach to overcome heterogeneities in identifiers inside queries. We present an ontology classifying the most common identifier semantic heterogeneities, and a service that makes use of it to cope with the problem using the described approach. Finally, we illustrate the solution by analysing a set of real queries.

Keywords. Ontologies, Clinical Trials, Database Integration, Semantic Mediation

Introduction

Advancing Clinico-Genomic Trials on Cancer (ACGT) is a European Commission-supported project that aims to develop a set of ontology-driven technologies to support the development of treatment and research within post-genomic multicentric clinical trials on cancer. One of the main issues that ACGT is trying to tackle is providing seamless, homogeneous syntactic and semantic access to a single virtual repository representing the set of integrated databases needed within a specific clinical trial. To achieve this goal, a data access infrastructure is being developed. This infrastructure is comprised by several tools: the ACGT Master Ontology on Cancer—an ontology modelling the domain of clinical trials on cancer—, the ACGT Data Access Services—coping with syntactic heterogeneities—and the ACGT Semantic Mediator—providing database integration and semantic homogenization. In order to facilitate the communication between different services, XML has been selected as the information representation format: the Master Ontology is developed in OWL, database schemas

are represented in RDFS by the Data Access Services and the adopted query language is SPARQL.

The ACGT Semantic Mediator deals with a variety of problems related to data semantic heterogeneities. These heterogeneities can be classified in two categories: i) Schema Level Heterogeneities, and ii) Instance/identifier Level heterogeneities. While the former are treated by mapping the RDFS representing the database schema information using the Master Ontology as a semantic framework, the latter cannot be tackled using this model. During the normal execution of a query through the Semantic Mediator, instance level heterogeneities are present in two different places: i) the retrieved results, and ii) identifiers in queries. We have proposed two approaches to address these issues. The first one is called OntoDataClean [1], and uses an ontology that represents the most common cases of instance level heterogeneity in the biomedical data. In this paper, we present the second approach, to overcome instance level heterogeneities in queries. It is based on the same principle proposed on OntoDataClean: using a transversal ontology representing the domain of instance level heterogeneities to produce a query translation overcoming these issues.

This paper is organized as follows: first, we give an overview of the state of the art and related work on semantic mediation and the study of heterogeneities. Then, we describe our approach and tool to overcome query identifier heterogeneities. In the case study section we show how our approach behaves with several real queries. Finally, we conclude summarizing our experiences so far.

1. Background & Related Work

The problem of database integration has been intensely studied during the last 20 years. We can classify the main approaches in three categories: i) Data Translation, ii) Query Translation, and iii) Information Linkage. Data Translation approaches are based in the actual homogenization of data for its subsequent storage in a central repository. Conversely, Query Translation approaches leave the data in the original databases, and expose a view to the user expressing the set of possible queries. When a query is launched, the mediation software splits and translates it into dedicated queries for the underlying databases. By contrast, Information Linkage approaches maintain a set of links among sources using cross reference. Both Data Translation and Query Translation approaches must take into consideration the possible semantic heterogeneities, but only the latter need to process queries in order to overcome them.

Formal semantics have formed part of the main efforts carried out in the fields of Knowledge Recovery and Database Management during the last years [2]. According to [3], XML can be used to overcome syntactic heterogeneities, while extensions of it, such as OWL introduce the semantic models as descriptions of domains—i.e. ontologies. However, semantic problems such as dealing with different descriptions of a single concept must be solved, even when ontologies are used as homogenization framework.

The classical role of ontologies in database integration is acting as a general description of the domain the different data belong to —e.g. clinical, genetic and image data, in the case of ACGT. Domain information from the ontology is mapped to terms and relations belonging to the databases. This mapping information will be used to translate a query or the actual data, depending on the selected approach. This kind of usage of ontologies has obtained success in different projects, such as ONTOFUSION

[4], SEMEDA [5], KAON Reverse [6] and D2RMAP [7]. These kinds of approaches treat mainly schema level heterogeneities.

By contrast, less amount of effort has been expended in resolving heterogeneities in the actual data—i.e. instance level heterogeneities. An example of tool that makes use of ontologies to overcome instance level heterogeneities in data is OntoDataClean [1]. This tool performs data processing in the results retrieving process, using an ontology that classifies the most common cases of semantic heterogeneity to identify and process data. To the best of our knowledge, no automatic ontology-based method to process identifier query heterogeneities exists. Some interesting studies on the types of identifier heterogeneities can be found, such as the classification given by [8].

2. Method

There exist different types of semantic heterogeneities that can be found when dealing with data coming from different sources, maybe belonging to heterogeneous domains. When formulating a query, a user can express the constraints of the view she wants to be retrieved by using explicit literals as values for determined fields within a database. A single database can present heterogeneity in its contained data—i.e. mainly because of bad information management before the storage, or the use of imprecise tools to collect them—, but the problems become greater when dealing with integrated databases. In the ideal case, a user that needs to launch a query does not want to be concerned about possible heterogeneities in the underlying data. Moreover, she may want to be given a standard way to express both terms and literals in her query—e.g. data can present scale representation heterogeneities, such as number of white blood cells per mililiter or microliter, depending on the laboratory that did the blood analysis. The question is the following: How a clinician should express a constraint in a query? The answer is simple: the system should give the user the view that there is only one way, which is the selected standard, and should be able to tackle deviations from this standard in the queries if they are presented. The mediation system must undertake then the responsibility of translating not only the terms, but the literals expressed in the query so that proper subqueries are sent to the underlying databases.

We propose an ontology-based approach, OntoQueryClean, to overcome these kinds of heterogeneities. OntoQueryClean is the evolution of OntoDataClean, a tool for data preprocessing also based on ontologies, and which forms part of the ONTOFUSION system for integration of distributed and heterogeneous data sources. OntoDataClean uses an ontology that proposes a classification of possible heterogeneities, and that defines a set of basic transformation methods. Instances of this preprocessing ontology can be created in order to specify how data have to be modified, allowing proper integration of heterogeneous sources. In OntoQueryClean, a new preprocessing ontology has been built, based on query preprocessing requirements, describing the most commong heterogeneities that can be found. Three transformation methods are defined, namely i) Scale transformations, ii) Format transformations and iii) Synonym transformations. Scale transformations allow solving heterogeneities due to the use of different scales among a set of sources—i.e. by defining arithmetic expressions that are applied over numeric data—, such as the previous example regarding measurements of white blood cells. Format transformations permit modifying string values through either regular expressions or rule algorithms. With these transformations we can, for example, modify the format of a date from MM-DD-

YYYY to DD/MM/YY. Synonym transformations allow defining pairs of synonyms so the system performs adequate substitutions.

Prior to actual query cleaning, the literals inside a specific query must be properly identified—i.e. recognize which concept they are instantiating. For this, a specific purpose module has been developed. This module is able to parse SPARQL queries and RDF Schema files to associate each literal included in the given query with a class in the schema. Figure 1 depicts the architecture of OntoQueryClean.

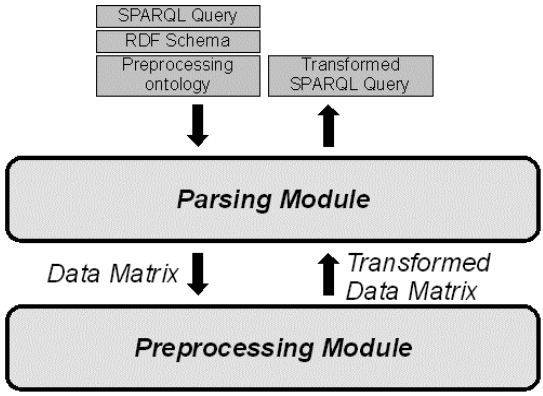

Figure 1. Architecture of the OntoQueryClean system

OntoQueryClean is part of the ACGT Semantic Mediator, taking care of the query preprocessing. Whenever a query is launched against the Mediator, this must create equivalent subqueries for the underlying sources. Once these are generated, they are given to the OntoQueryClean tool so literals are properly transformed and follow the format imposed by each database. Previous work has to be carried out in order to define preprocessing ontologies for each source. This task can be done using any existing ontology editor, such as Protégé or SWOOP.

3. Results

We have tested our tools by conducting experiments using three different data sources, from two clinical trials ACGT is working in—i.e. TOP and SIOP—and a DICOM repository of images. Test versions of these databases were incorporated in the ACGT Semantic Mediator infrastructure, so integrated queries could be performed. As stated before, OntoQueryClean is accessed once the schema level heterogeneities have been already tackled by the Mediator. A set of integrated queries was launched against the Mediator, and thus preprocessed by OntoQueryClean. In the following paragraphs we describe two of these queries, showing how the tool solved the heterogeneities found.

The first example shows a synonym transformation applied over a literal indicating a patient identifier. A global query asking for hospital identifiers where a specific patient was treated and clinical study identifiers for that same patient is presented. This query involves only the SIOP and DICOM databases, thus only queries for these are generated. The identifier used in the global query corresponds to the existing one in SIOP. However, it differs from the identifiers contained in DICOM. Synonym

transformations are carried out in order to convert the string identifiers into the format used in DICOM. Figure 2 shows the queries before and after translation.

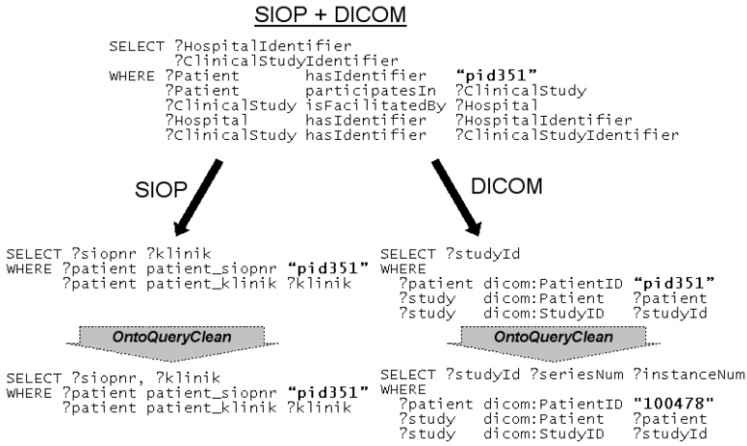

Figure 2. Synonym transformations over an integrated query for SIOP + DICOM

As can be seen, a general query expressed in terms of the global schema of the integrated repository is launched. This query includes a constraint on the patient identifier. The Mediator undertakes the task of splitting this query into dedicated subqueries for both SIOP and DICOM. However, the Mediator itself does not deal with the possible heterogeneity in the literal included in the constraint. In order to cope with this issue, OntoQueryClean is invoked. In the case of the SIOP query, OntoQueryClean finds no differences between the literal format of the query produced by the Mediator and the one to be sent to the underlying database. By contrast, in the case of DICOM, a different format is used to instantiate patient identifiers. OntoQueryClean translates the original identifier into the proper value for DICOM.

The second example involved a format transformation applied over a different integrated query. In this case, the mapping for this query produced subqueries only for TOP and SIOP. A restriction on the date of a specific treatment on patients was included in the original query. In the global schema, values for dates are separated in three different fields—i.e. day, month and year—but both TOP and SIOP employ a single field with the union of the mentioned fields—with the format DD-MM-YYYY—. We used the format transformation method of OntoDataClean to overcome this heterogeneity, successfully producing the expected literals in the restrictions.

It must be also noted that, even though no exact measures were performed, no significant increase in latency of the system was noticed when including the OntoQueryClean tool in the Mediator.

4. Conclusions

We have described a method and developed a tool for query heterogeneities preprocessing. The method presented is based on the use of ontologies to classify the most common query identifier heterogeneities, and to automate the query identifier transformation task. A dedicated ontology describing the domain has been described, based on previous works on this task. Initially, three transformations methods are

described—namely: scale, format and synonym transformations. The ontology can be extended easily to support future requirements.

The current implementation of OntoQueryClean tool supports the described methods, which cover the most important problems when dealing with literal heterogeneity in query translation. This tool allows transformation of literals inside SPARQL queries, facilitating the integration of heterogeneous sources. OntoQueryClean has been successfully integrated into the ACGT Semantic Mediator, a mediation system being implemented inside the ACGT project. The tool is composed of two modules: i) the parsing module and ii) the preprocessing module. The former is devoted to parsing SPARQL queries in order to extract the literals that must be transformed, while the latter performs the actual transformations.

Two experiments were presented, based on real queries formulated by clinicians, proving the suitability of our approach. In both cases, the tool successfully generated the expected results, allowing to properly querying the respective data sources.

Future work will involve increasing the features offered by the tool by embracing more transformation methods and by offering support to different query languages. We plan to include machine methods to automate the task of defining the transformation, such as the use of similarity functions. These improvements will lead to a more general purpose tool, able to cope with the requirements imposed by a wider set of domains.

5. Acknowledgements

This work was funded by the European Commission, through the ACGT integrated project (FP6-2005-IST-026996).

References

[1] D. Perez-Rey, A. Anguita, and J. Crespo: OntoDataClean: Ontology-based Integration and Preprocessing of Distributed Data, *Lecture notes in Computer Science* **4345** (2006), 262-272.

[2] N. Noy: Order from Chaos, *ACM Queue vol. 3* **8** (2005). Available at: http://www.acmqueue.org/modules.php?name=Content&pa=showpage&pid=341&page=1

[3] A. Halevy: Why Your Data Won't Mix, *ACM Queue vol. 3* **8** (2005). Available at: http://www.acmqueue.org/modules.php?name=Content&pa=showpage&pid=336

[4] D. Perez-Rey, V. Maojo, M. Garcia-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Sanchez, and A. Sousa: ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases, *Computers in Biology and Medicine* **36** (2006), 712-730.

[5] J. Köhler, S. Philippi, and M. Lange: SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics vol 19* 18 (2003), 2420-2427.

[6] G.R. Librelotto, W. Souza, J.C. Armalo, and P.R. Henriques, Using the Ontology Paradigm to Integrate Information Systems, *Proceedings of the International Conference on Knowledge Engineering and Decision Support,* Porto (2004), 497–504.

[7] C. Bizer: D2R MAP - A Database to RDF Mapping Language, *Proceedings of the International World Wide Web Conference,* Budapest (2003)

[8] Sources and Classification of Semantic Heterogeneities. Available at: http://www.mkbergman.com/?p=232