Pontifícia Universidade Católica
do Rio de Janeiro

Luiz André Portes Paes Leme

Conceptual schema matching based on similarity heuristics

D.Sc. Thesis

Thesis presented to the Graduate Program of the Departamento de Informática, PUC–Rio in partial fulfillment of the requirements for the degree of Doctor of Science in the subject of Informatics.

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro
March 23, 2009

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Luiz André Portes Paes Leme

Conceptual schema matching based on similarity heuristics

Thesis presented to the Graduate Program of the Departamento de Informática, PUC–Rio in partial fulfillment of the requirements for the degree of Doctor of Science in the subject of Informatics.

Prof. Marco Antonio Casanova
Advisor
Departamento de Informática – PUC-Rio

Prof. Antonio L. Furtado
Departamento de Informática – PUC-Rio

Prof. Ruy Luiz Milidiú
Departamento de Informática – PUC-Rio

Profª. Karin Koogan Breitman
Departamento de Informática – PUC-Rio

Profª. Cláudia Bauzer Medeiros
Instituto de Computação – UNICAMP

Prof. Angelo Ernani Maia Ciarlini
Departamento de Informática Aplicada – UNIRIO

Prof. José Eugenio Leal
Coordenador Setorial de Pós-Graduação do CTC – PUC-Rio

Rio de Janeiro – March 23, 2009

Luiz André Portes Paes Leme

graduated in Electrical Engineering at the Universidade do Estado do Rio de Janeiro (1989), has a Specialization Degree in Distributed Systems from the Núcleo de Computação Eletrônica da UFRJ (2001), and received his Master Degree in Informatics at the Pontifical Catholic University of Rio de Janeiro (2006). He has been acting in software engineering and database design since 1989 for consulting and health plan companies.

Ficha Catalográfica

Leme, Luiz André Portes Paes

Conceptual schema matching based on similarity heuristics / Luiz André Portes Paes Leme ; orientador: Marco Antonio Casanova. – 2009.

106 f. : il. ; 30 cm

Tese (Doutorado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Inclui bibliografia

1. Informática – Teses. 2. Banco de dados. 3. Alinhamento de esquema conceitual. 4. Similaridade. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my wife, because of her love and patience in helping me overcome the difficulties.

Acknowledgments

To my advisor Prof. Marco Antonio Casanova for the great stimulus and technical knowledge.

To my colleagues at PUC-Rio for innumerous contributions to this work.

To CNPq and PUC-Rio, for the grants which made possible this work.

To the professors which took part in the examination board for their contributions.

To all professors and employees of Informatics Department of PUC-Rio for their precious technical knowledge and help.

To all my family and friends that have encouraged or helped me in any way.

Resumo

Leme, Luiz André Portes Paes; Casanova, Marco Antonio. **Alinhamento de esquemas conceituais baseado em heurísticas de similaridade.** Rio de Janeiro, 2009. 106p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Alinhamento de esquema é uma questão fundamental em aplicações de banco de dados, tais como mediação de consultas, integração de banco de dados e armazéns de dados. Nesta tese, abordamos inicialmente o alinhamento de catálogos. Um catálogo é um banco de dados simples que contém informações sobre conjuntos de objetos, tipicamente classificados usando-se termos de um dado tesauro. Inicialmente apresentamos uma técnica de alinhamento baseada na noção de similaridade, que se aplica a pares de tesauros e de listas de propriedades. Descrevemos, então, o alinhamento baseado na noção de informação mútua e introduzimos variações que exploram certas heurísticas. Ao final, discutimos resultados experimentais que avaliam a precisão do método e que comparam a influência das heurísticas. Após as técnicas para alinhamento de catálogos, nos concentramos no problema mais complexo de alinhamento de dois esquemas descritos em um subconjunto de OWL. Adotamos uma técnica baseada em instâncias e, por isso, assumimos que conjuntos de instâncias de cada esquema estão disponíveis. Decompomos este problema nos subproblemas de alinhamento de vocabulário e de alinhamento de conceitos. Introduzimos também condições suficientes para garantir que o alinhamento de vocabulário induz um alinhamento de conceitos correto. Em seguida, descrevemos uma técnica de alinhamento de esquemas OWL baseada no conceito de similaridade. Finalmente, avaliamos a precisão da técnica usando dados disponíveis na Web. De forma diferente de outras técnicas anteriores baseadas em instâncias, o processo de alinhamento que descrevemos usa funções de similaridade para induzir alinhamento de vocabulários de uma forma não trivial. Ilustramos, também, que a estrutura de esquemas OWL pode nos levar a mapeamentos de conceitos errados e indicamos como evitar tais problemas.

Palavras-chave

Banco de dados. Alinhamento de esquemas conceituais. Similaridade.

Abstract

Leme, Luiz André Portes Paes; Casanova, Marco Antonio. **Conceptual schema matching based on similarity heuristics.** Rio de Janeiro, 2009. 106p. DSc Thesis – Department of Informatics – Pontifical Catholic University of Rio de Janeiro.

Schema matching is a fundamental issue in many database applications, such as query mediation, database integration, catalog matching and data warehousing. In this thesis, we first address how to match catalogue schemas. A catalogue is a simple database that holds information about a set of objects, typically classified using terms taken from a given thesaurus. We introduce a matching approach, based on the notion of similarity, which applies to pairs of thesauri and to pairs of lists of properties. We then describe matchings based on cooccurrence of information and introduce variations that explore certain heuristics. Lastly, we discuss experimental results that evaluate the precision of the matchings introduced and that measure the influence of the heuristics. We then focus on the more complex problem of matching two schemas that belong to an expressive OWL dialect. We adopt an instance-based approach and, therefore, assume that a set of instances from each schema is available. We first decompose the problem of OWL schema matching into the problem of vocabulary matching and the problem of concept mapping. We also introduce sufficient conditions guaranteeing that a vocabulary matching induces a correct concept mapping. Next, we describe an OWL schema matching technique based on the notion of similarity. Lastly, we evaluate the precision of the technique using data available on the Web. Unlike any of the previous instance-based techniques, the matching process we describe uses similarity functions to induce vocabulary matchings in a non-trivial way, coping with an expressive OWL dialect. We also illustrate, through a set of examples, that the structure of OWL schemas may lead to incorrect concept mappings and indicate how to avoid such pitfalls.

Keywords

Database. Conceptual schema matching. Similarity.

Contents

List of figures

List of tables

List of Symbols

S – source schema

s – element of the schema S

T – target schema

t – element of schema T

R – set of RDF triples that defines the properties and classes of a database schema

o_R – subset of an extension R, usually a sample of the database defined by R

$o_R[P]$ – values of the property P in o_R

$o_R[C]$ – instance ids (URIrefs) of the Class C of R in o_R

$R[A_1, A_2, \ldots, A_n]$ – relation R with properties (attributes) A_1, A_2, A_n

σ_R – set of tuples of $R[A_1, A_2, \ldots, A_n]$

$\sigma_R[A_i]$ – set of values of the property A_i in σ_R

$C[A_1\ U_1, ..., A_m\ U_m]$ – catalogue schema C where A_1 is the property which holds the id's of the catalogue entries, A_2 is the classification of the entries, $A_3, \ldots, A_n$ are the properties of the entries and $U_1, \ldots, U_n$ are the ranges of each property.

U_C – set of RDF triples that defines instances of C, a sample of the database

I, J – instances of catalogues

μ_P – property matching between two catalogue schemas

μ_T – thesauri terms matching between two catalogue schemas

μ_I – instance matching between two schemas

$o[C,A_i]$ – values of the property A_i in U_C

$ot[C,A_i]$ – tokens of a string property A_i in U_C

$iv[C,A_i]$ – *(instance id,value)* ordered pairs of the property A_i in U_C

$i[C,t]$ – set of instance id's classified as t in a thesaurus T

V_S – set of properties and classes of the database schema S

C_S – set of classes in V_S

P_S – set of properties in V_S

μ_V – vocabulary matching between two database schemas

μ_C – class matching between two database schemas

props[S,C] – properties of the class *C* in the database schema *S*

relprops[S,C,T,D] – properties of class C in the database schema S where the matching properties with the class D of schema T is replaced by the equivalent property of D.

P^C – contextualized property *P*: a property *P* attached to instances of the class *C*

tp – amount of true positive matchings. A true positive matching is that matching either pointed out by the matching algorithm and that it comes to be valid given a matching reference.

fp – amount of false positive matchings. A false positive matching is that matching pointed out by the matching algorithm, but that it is not valid given a matching reference.

tn – the amount of true negatives matchings. A false negative matching is that matching either missed by the matching algorithm and that it is not valid given a matching reference.

fn – the amount of false negatives matchings. A false negative matching is that matching missed by the matching algorithm, but it comes to be true given a matching reference.