Pontifícia Universidade Católica
do Rio de Janeiro

Daniela Francisco Brauner

Alinhamento de esquemas baseado em instâncias

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro
Junho de 2008

Pontifícia Universidade Católica
do Rio de Janeiro

Daniela Francisco Brauner

Alinhamento de esquemas baseado em instâncias

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marco Antonio Casanova
Departamento de Informática - PUC-Rio (Orientador)

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

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

Prof. Carlos José Pereira de Lucena
Departamento de Informática – PUC-Rio

Profª. Cláudia Bauzer Medeiros
Instituto de Computação - UNICAMP

Profª. Vânia Maria Ponte Vidal
Departamento de Computação - UFC

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 09 de junho de 2008

Daniela Francisco Brauner

Mestre em Informática pela PUC-Rio em abril de 2005. Graduou-se em Ciência da Computação pela UFPel (Universidade Federal de Pelotas) em março de 2003.

Ficha Catalográfica

Brauner, Daniela Francisco

Alinhamento de esquemas baseado em instâncias / Daniela Francisco Brauner ; orientador: Marco Antonio Casanova. – 2008.

83 f. : il. ; 30 cm

Tese (Doutorado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Alinhamento de esquema. 3. Esquema conceitual. 4. Tesauro. 5. Banco de dados. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Em primeiro lugar, quero agradecer ao meu orientador, Prof. Casanova, pela confiança que depositou em mim, pelas críticas e sugestões que fez ao longo deste trabalho e, principalmente, pelo apoio constante.

Um agradecimento especial ao Prof. Carlos Lucena pela confiança, orientação e pelas oportunidades concedidas. Agradeço também a alocação no Laboratório de Engenharia de Software, onde encontrei um espaço privilegiado de criatividade, dinamismo e amizade.

Agradeço ao Programa de Pós-Graduação em Informática da PUC-Rio que me permitiu usufruir de sua excelência em ensino e pesquisa, com especial agradecimento ao corpo docente que, por meio de seus cursos, me permitiu aprimorar e ampliar meus conhecimentos na área para desenvolver esta pesquisa.

Agradeço ao CNPq pelo financiamento.

Aos meus amigos, por entenderem a minha ausência e apoiarem as minhas decisões. Em especial à minha grande amiga Vera, por seu exemplo de vida e seus conselhos memoráveis.

Aos meus pais, por me incentivarem a enfrentar os desafios.

Ao meu noivo, Alexandre, por todo carinho, compreensão e principalmente, pela companhia incondicional.

Finalmente, agradeço à Deus!

Resumo

Brauner, Daniela Francisco; Casanova, Marco Antonio. **Alinhamento de esquemas baseado em instâncias.** Rio de Janeiro, 2008. 83p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Um mediador é um componente de software que auxilia o acesso a fontes de dados. Com o advento da Web, a construção de mediadores impõe desafios importantes, tais como a capacidade de fornecer acesso integrado a fontes de dados independentes e dinâmicas e a habilidade de resolver a heterogeneidade semântica entre os esquemas destas fontes. Para lidar com esses desafios, o alinhamento de esquemas é uma questão fundamental. Nesta tese são propostas abordagens de alinhamento de esquemas de classificação (tesauros) e esquemas conceituais, utilizando instâncias como evidências para os mapeamentos. As abordagens propostas são classificadas em dois tipos: adaptativa e *a priori*, referindo-se, respectivamente, à descoberta dos mapeamentos de forma incremental ou à definição dos mapeamentos antes da implantação do mediador. Por fim, são apresentados experimentos para validação e teste das abordagens propostas.

Palavras-chave

Alinhamento de esquemas, esquema conceitual, tesauro, banco de dados.

Abstract

Brauner, Daniela Francisco; Casanova, Marco Antonio. **Instance-based Schema Matching.** Rio de Janeiro, 2008. 83p. DSc. Thesis - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A mediator is a software component that helps accessing data sources. With the advent of the Web, the design of mediators imposes important challenges, such as the ability of providing integrated access to independent and dynamic data sources and the ability of resolving the semantic heterogeneity between different data source schemas. To deal with these challenges, schema matching is a fundamental issue. In this thesis, matching approaches for classification schemas (thesauri) and conceptual schemas are proposed, using instances as evidences for the mappings. The proposed approaches are classified as adaptative and *a priori*, referring to, respectively, the discovery of the mappings in an incremental way or the definition of the mappings before the deployment of the mediator. Finally, experiments to validate and test the proposed approaches are presented.

Keywords

Schema matching, conceptual schema, thesaurus, database.

Sumário

Lista de figuras

Lista de tabelas

Lista de siglas

ADL	Alexandria Digital Library
ADL FTT	Alexandria Digital Library Feature Type Thesaurus
DW	Data Warehouse
ER	Entidade-Relacionamento
ESRB	Entertainment Software Rating Board
ETL	Extract - Transformation - Load
GEOnet	GEOnet Names Server
GeoNames	GeoNames geographical database
GTIN	Global Trade Item Number
HTML	HyperText Markup Language
ISBN	International Standard Book Number
ISO	International Standards Organization
MAFL	Módulo de Acesso às Fontes Locais
MAFR	Módulo de Acesso às Fontes Remotas
MEM	Módulo de Esquema Mediado
METM	Módulo Estimador de Taxas de Mapeamento
MGC	Módulo Gerente de Consultas
MI	Módulo de Interface
MR	Módulo de Registro
NGA	National Geospatial Intelligence Agency
OWL	Ontology Web Language
RDF	Resource Description Framework
SGBD	Sistema Gerenciador de Bancos de Dados
SGM	Sistema de Gerenciamento de Modelos
SQL	Structured Query Language
TI	Tecnologia da Informação
UPC	Universal Product Code
W3C	World Wide Web Consortium
XML	eXtensible Markup Language