Pontifícia Universidade Católica
do Rio de Janeiro

Alexandre Gazola

Uma infra-estrutura de software para alinhamento de catálogos heterogêneos

Dissertação de Mestrado

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

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro

Março de 2008

Pontifícia Universidade Católica
do Rio de Janeiro

Alexandre Gazola

Uma infra-estrutura de software para alinhamento de catálogos heterogêneos

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

Prof. Marco Antonio Casanova
Orientador
PUC-Rio

Prof. Antonio Luz Furtado
PUC-Rio

Profª. Karin Koogan Breitman
PUC-Rio

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

Rio de Janeiro, 27 de março de 2008

Alexandre Gazola

Graduou-se em Ciência da Computação pela Universidade Federal de Viçosa (UFV) em maio de 2006. Em 2007 foi premiado pelo programa Bolsa Nota 10 da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), destinado aos melhores alunos dos melhores Programas de Pós-Graduação do Estado do Rio de Janeiro. Tem experiência na área de Ciência da Computação, com ênfase em Banco de Dados e Engenharia de Software. É editor técnico da revista Mundo Java, publicação nacional sobre desenvolvimento de software e tecnologia Java.

Ficha Catalográfica

Gazola, Alexandre

Uma infra-estrutura de software para alinhamento de catálogos heterogêneos / Alexandre Gazola ; orientador: Marco Antonio Casanova. – 2008.

87 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Catálogo de objetos. 3. Esquema. 4. Tesauro. 5. Instância. 6. Alinhamento. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

A Jesus Cristo, meu melhor amigo

Agradecimentos

Agradeço primeiramente a Deus, pelo seu imensurável amor derramado na cruz por meio de seu Filho Jesus Cristo. Tudo que sou e tudo que conquistei até hoje provém desse Deus maravilhoso, que sabe dar boas dádivas a seus filhos. Ele é digno de “receber o poder, e riqueza, e sabedoria, e força, e honra, e glória, e louvor”. (Ap 5:12)

Agradeço à minha querida família, pelo conforto e fundamental apoio em todo tempo. Agradeço ao meu pai Nelson, minha mãe Heloísa, meu irmão Rafael e minha irmã Thalita. À minha noiva, Samira, por ter estado sempre ao meu lado, trazendo motivação e carinho, apesar das dificuldades.

Ao meu orientador, o professor Marco Antonio Casanova, pela confiança depositada em mim, por seu auxílio durante este trabalho, pelas oportunidades que me concedeu e por sua paciência.

Aos colegas de curso Algemiro, Fábio Guerra, André Marins e Ricardo Leal. Deixo um agradecimento especial à Daniela Brauner, por ter contribuído de maneira fundamental para que este trabalho se concretizasse.

A todos os colegas do laboratório TecGraf, principalmente ao Vinci Amorim, Roberto Santos, Demétrius Nunes, Eduardo David, Luis Gustavo Ferrão, Luiz Gustavo Ferreira da Silva Costa, Sandro Marinho e Viliam.

À Patrícia, Adriana, Fábio e dona Maria pelo importante suporte que me deram para que eu pudesse me estabelecer no Rio de Janeiro; e à Cecília, por ter me aceitado em sua casa e me proporcionado um ótimo lugar para residir durante meus estudos.

À UFV e ao professor Jugurta Lisboa Filho, pelo aprendizado e pelas oportunidades concedidas.

À CAPES, e à FAPERJ pelo financiamento desta pesquisa.

Resumo

Gazola, Alexandre; Casanova, Marco Antonio. **Uma infra-estrutura de software para alinhamento de catálogos heterogêneos.** Rio de Janeiro, 2008. 87p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A maior parte dos bancos de dados existentes é projetada de maneira independente e, portanto, é geralmente implementada utilizando diferentes esquemas conceituais, criando um contexto de heterogeneidade em níveis sintático, estrutural e semântico. Não obstante, quando um conjunto de bancos de dados se refere a um mesmo domínio, eventualmente, surge a necessidade de integrá-los em um mesmo banco, ou de intermediar o acesso ao conjunto de bancos de forma transparente. Para tratar o problema da heterogeneidade, torna-se necessário o alinhamento dos esquemas de cada um dos bancos de dados envolvidos. Esse processo geralmente é feito por especialistas de domínio, mas tende a ser um trabalho muito tedioso e propenso a erros. Esta dissertação apresenta o *CatalogMatcher*, uma infra-estrutura de software para alinhamento de catálogos heterogêneos. Um catálogo armazena dados sobre um conjunto de objetos de um determinado domínio, tipicamente classificados por algum tipo de taxonomia ou tesauro. O *CatalogMatcher* contém componentes que implementam estratégias de alinhamento de catálogos heterogêneos utilizando abordagens baseadas em instâncias.

Palavras-chave

Catálogo de objetos; Esquema; Tesauro; Instância; Alinhamento.

Abstract

Gazola, Alexandre; Casanova, Marco Antonio. **A software infrastructure for catalog matching.** Rio de Janeiro, 2008. 87p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Most databases are independently designed and, therefore, are usually implemented using different conceptual schemas, which creates a context of syntactic, structural and semantic-level heterogeneity. Nevertheless, when a set of databases refers to a common domain, it may become necessary to integrate them into a single database, or to intermediate access to the databases in a transparent way. To deal with the heterogeneity problem, it becomes necessary to align the conceptual schemas. This process is usually carried out by domain specialists, and tends to be tedious and error-prone. This dissertation presents the *CatalogMatcher*, a software infrastructure for catalog matching. A catalog stores data about a set of objects from a specific domain, typically classified by some sort of taxonomy or thesaurus. The *CatalogMatcher* contains components that implement instance-based alignment strategies.

Keywords

Objects catalog; Schema; Thesaurus; Instance; Matching.

Sumário

Lista de figuras

Lista de tabelas