Pontifícia Universidade Católica
do Rio de Janeiro

Gustavo Lopes Mourad

Um Framework para a Construção de Mediadores Oferecendo Eliminação de Duplicatas

Dissertação de Mestrado

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

Orientador: Prof. Karin Breitman

Rio de Janeiro

Setembro de 2010

Pontifícia Universidade Católica
do Rio de Janeiro

Gustavo Lopes Mourad

Um Framework para a Construção de Mediadores Oferecendo Eliminação de Duplicatas

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Karin Breitman
Orientador
Departamento de Informática – PUC-Rio

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

Prof. Simone Diniz Junqueira Barbosa
Departamento de Informática – PUC-Rio

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

Rio de Janeiro, 14 de Setembro de 2010

Gustavo Lopes Mourad

Graduou-se em Engenharia de Computação pela Pontifícia Universidade Católica do Rio de Janeiro. Desenvolveu como projeto de mestrado um framework para enriquecimento de bases de dados com recursos da Deep Web.

Ficha Catalográfica

Mourad, Gustavo L.

Um Framework para a Construção de Mediadores Oferecendo Eliminação de Duplicatas / Gustavo Lopes Mourad; orientador: Karin Breitman. – Rio de Janeiro : PUC-Rio, Departamento de Informática, 2010

v., 69 f: il. ; 29,7 cm

1. Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Dissertação. 2. Detecção de duplicatas. 3. Resolução de entidades . 4. Integração de Dados 5. Deep Web. I. Breitman, Karin. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título

CDD: 004

Dedico este trabalho a meus pais e avós pelo apoio aos estudos em todas as fases da minha vida, minha esposa Rachel e minha irmã.

Agradecimentos

A minha orientadora, Dr. Karin Breitman, por ter me aceito e acreditado no meu trabalho. Sua competência, dedicação e paciência foram fundamentais para a realização desta dissertação.

Aos professores do Departamento de Informática da PUC-Rio pelas aulas, conselhos e questões estimulantes.

A todos os colegas de curso e trabalho que sempre me apoiaram com coleguismo, respeito e amizade.

Por fim, agradeço à CAPES e PUC-Rio pelo apoio financeiro e ao Departamento de Informática da PUC-Rio pela excelente formação.

Resumo

Mourad, Gustavo L.; Breitman, K. **Um Framework para Construção de Mediadores Oferecendo Eliminação de Duplicatas**. Rio de Janeiro, 2010. 69p**.** Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

À medida em que aplicações web que combinam dados de diferentes fontes ganham importância, soluções para a detecção online de dados duplicados tornam-se centrais. A maioria das técnicas existentes são baseadas em algoritmos de aprendizado de máquina, que dependem do uso de bases de treino criadas manualmente. Estas soluções não são adequadas no caso da Deep Web onde, de modo geral, existe pouca informação acerca do tamanho das fontes de dados, da volatilidade dos mesmos e do fato de que a obtenção de um conjunto de dados relevante para o treinamento é uma tarefa difícil. Nesta dissertação propomos uma estratégia para extração (scraping), detecção de duplicatas e incorporação de dados resultantes de consultas realizadas em bancos de dados na Deep Web. Nossa abordagem não requer o uso de conjuntos de testes previamente definidos, mas utiliza uma combinação de um classificador baseado no Vector Space Model, com funções de cálculo de similaridade para prover uma solução viável. Para ilustrar nossa proposta, nós apresentamos um estudo de caso onde o framework é instanciado para uma aplicação do domínio dos vinhos.

Palavras-chave

Detecção de Duplicatas; Resolução de Entidades; Integração de dados; Deep Web

Abstract

Mourad, Gustavo L.; Breitman, K. (advisor). **A Framework for the Construction of Mediators Offering Deduplication**. Rio de Janeiro, 2010. 69p**.** MSc Dissertation - Departmento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As Web applications that obtain data from different sources (Mashups) grow in importance, timely solutions to the duplicate detection problem become central. Most existing techniques, however, are based on machine learning algorithms, that heavily rely on the use of relevant, manually labeled, training datasets. Such solutions are not adequate when talking about data sources on the Deep Web, as there is often little information regarding the size, volatility and hardly any access to relevant samples to be used for training. In this thesis we propose a strategy to aid in the extraction (scraping), duplicate detection and integration of data that resulted from querying Deep Web resources. Our approach does not require the use of pre-defined training sets, but rather uses a combination of a Vector Space Model classifier with similarity functions, in order to provide a viable solution. To illustrate our approach, we present a case study where the proposed framework was instantiated for an application in the wine industry domain.

Keywords

Duplicate Detection; Entity Matching; Data Integration; Deep Web

Sumário

Lista de Figuras

True genious resides in the capacity for evaluation of uncertain, hazardous, and conflicting information.

Winston Churchill