



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



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

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

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 - Departamento 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

| | | |
|-----|---|----|
| 1 | Introdução | 12 |
| 1.1 | Objetivo | 13 |
| 1.2 | Contribuições | 14 |
| 1.3 | Resumo | 15 |
| 2 | Conceitos Básicos | 16 |
| 2.1 | Recuperação de Dados da Web | 16 |
| 2.2 | Construção de rastreadores (<i>crawlers</i>) | 18 |
| 2.3 | Alinhamento de Esquemas - Schema Matching | 20 |
| 2.4 | Resolução de entidades – Entity Matching | 20 |
| 2.5 | Funções de Similaridade entre Cadeias de Caracteres | 24 |
| 2.6 | Classificação | 29 |
| 2.7 | Resumo | 32 |
| 3 | Estratégia para o enriquecimento de informações | 34 |
| 3.1 | Busca de informações na Deep Web | 35 |
| 3.2 | Incorporação de novas informações a partir da identificação de duplicatas | 36 |
| 3.3 | Escopo do trabalho | 39 |
| 3.4 | Resumo | 40 |
| 4 | Framework Proposto para Construção de Mediadores | 41 |
| 4.1 | Visão Geral do Processo do Framework Proposto | 41 |
| 4.2 | Arquitetura Proposta | 44 |
| 4.3 | Resumo | 48 |
| 5 | Winetag.com.br: um estudo de caso | 49 |
| 5.1 | Metodologia de trabalho | 50 |
| 5.2 | Resultados | 58 |
| 5.3 | Resumo | 62 |
| 6 | Trabalhos relacionados | 63 |
| 6.1 | Rastreadores para Deep Web | 63 |
| 6.2 | Resolução de Entidades | 63 |
| 6.3 | Comparação de resultados | 64 |

| | | |
|-----|----------------------------|----|
| 7 | Conclusão | 65 |
| 7.1 | Contribuições | 65 |
| 7.2 | Trabalhos Futuros | 66 |
| 8 | Referências Bibliográficas | 67 |

Lista de Figuras

| | |
|---|----|
| Figura 1 - Vista conceitual da Deep Web [17] | 12 |
| Figura 2 - Classificação de Extratores Web [23] | 17 |
| Figura 3 - Rastreador (<i>crawler</i>) tradicional [32] | 18 |
| Figura 4 - Rastreador (<i>crawler</i>) de Deep Web [32] | 19 |
| Figura 5 - Visão geral do algoritmo de similaridade de MARLIN [6] | 23 |
| Figura 6 - Representação gráfica de 2 documentos no VSM | 27 |
| Figura 7 - Exemplo de execução do KNN | 30 |
| Figura 8 - Exemplo de árvore de decisão | 31 |
| Figura 9 - Exemplo de estrutura básica de nós de uma rede neural | 32 |
| Figura 10 - Visão geral do processo de enriquecimento de informações proposto: (a) Busca, (b) Incorporação de dados | 34 |
| Figura 11 - Algoritmo para a busca de informações na Deep Web | 36 |
| Figura 12 - Algoritmo para o resolução de entidades | 38 |
| Figura 13 - Passos 1, 2 (a) e 3 (b) do processo (estratégia de busca) | 42 |
| Figura 14 - Passos 4 e 5 do processo (incorporação) | 43 |
| Figura 15 - Diagrama de Classes | 45 |
| Figura 16 - Diagrama de Componentes do Framework | 48 |
| Figura 17 - Exemplo de fonte de dados | 53 |
| Figura 18 - Exemplo elementos identificados pelo Developer Tools | 53 |
| Figura 19 - Elementos destacados no código-fonte | 54 |
| Figura 20 - Diagrama de classes para instância de vinhos | 57 |
| Figura 21 - Winetag.com.br: mashup com dados da Deep Web | 58 |

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

Winston Churchill