Pontifícia Universidade Católica
do Rio de Janeiro

Thiago Ribeiro Nunes

A Model for Exploration of Semi-Structured Datasets

TESE DE DOUTORADO

Thesis presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Doutor em Ciências - Informática

Advisor: Prof. Daniel Schwabe

Rio de Janeiro
October 2017

Pontifícia Universidade Católica
do Rio de Janeiro

Thiago Ribeiro Nunes

A Model for Exploration of Semi-Structured Datasets

Thesis presented to the Programa de Pós-Graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática. Approved by the undersigned Examination Committee.

Prof. Daniel Schwabe
Advisor
Departamento de Informática – PUC-Rio

Prof[a]. Adriana Pereira de Medeiros
Universidade Federal Fluminense – UFF

Prof. Edward Hermann Haeusler
Departamento de Informática – PUC-Rio

Prof. Hélio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Prof[a]. Maria Luiza Machado Campos
Universidade Federal do Rio de Janeiro – UFRJ

Prof. Sergio Lifschitz
Departamento de Informática – PUC-Rio

Prof. Márcio da Silveira Carvalho
Vice Dean of Graduate Studies
Centro Técnico Científico da PUC-Rio

Rio de Janeiro, October 6th, 2017

Thiago Ribeiro Nunes

Graduated in Computer Science from Universidade Cândido Mendes – UCAM in 2009, he obtained the degree of Doutor at PUC-Rio in 2017 in Ciências – Informática in 2017

Bibliographic data

Nunes, Thiago Ribeiro

A model for exploration of semi-structured datasets / Thiago Ribeiro Nunes ; advisor: Daniel Schwabe. – 2017.
144 f. : il. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2017.
Inclui bibliografia

1. Informática – Teses. 2. Exploração. 3. Modelo formal. 4. Framework. 5. Dados semiestruturados. I. Schwabe, Daniel. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

To God and Jesus Christ: my endless source of life inspiration, piece, and love.

I would like to express my sincere gratitude to my advisor Prof. Daniel Schwabe for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, friendship, and all knowledge and experience shared. His advices helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank my thesis committee: Profª. Adriana Pereira de Medeiros, Prof. Antonio Luiz Furtado, Prof. Edward Hermann Haeusler, Prof. Hélio Côrtes Vieira Lopes, Profª. Maria Luiza Machado Campos, Prof. Mark Douglas de Azevedo Jacyntho, and Prof. Sergio Lifschitz, for their availability and insightful comments.

I would like to thank all Departamento de Informática (DI) staff, for the support along this journey.

To Instituto Federal Fluminense, CAPES, CNPq, and Google Brazil Research Program for the financial support along this research project.

To my colleagues at TecWeb Lab, especially to Guilherme Szundy, for the technical support along this research.

To my wife Camila Botelho da Silva, for all emotional support, love, and motivation she gave me along this research, essential for my success.

To my parents Iva Rangel Ribeiro Nunes and Reginaldo Soares Nunes, for the unconditional love, education, values, motivation, and support they gave me throughout my life.

To my family, for the encouraging words, kindness, and emotional support in the difficult moments.

To my friends: Chrystiano Araújo, Marcelo Arruda, Leandro Cruz, José Eduardo Talavera Herrera, and Ana Laura Tavares Rodrigues, for the nice moments, encouraging words, and motivation.

And finally, last but by no means least, to my friends at Instituto Federal Fluminense campus Campos-Guarus for the support along this research.

Abstract

Nunes, Thiago Ribeiro; Schwabe, Daniel (Advisor). **A Model for Exploration of Semi-Structured Datasets**, Rio de Janeiro, 2017. 144p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Information exploration processes are usually recognized by their inherent complexity, lack of knowledge and uncertainty, concerning both the domain and the solution strategies. Even though there has been much work on the development of computational systems supporting exploration tasks, such as faceted search and set-oriented interfaces, the lack of a formal understanding of the exploration process and the absence of a proper separation of concerns approach in the design phase is the cause of many expressivity issues and serious limitations. This work proposes a novel design approach of exploration tools based on a formal framework for representing exploration actions and processes. Moreover, we present a new exploration system that generalizes the majority of the state-of-the art exploration tools. The evaluation of the proposed framework is guided by case studies and comparisons with state-of-the-art tools. The results show the relevance of our approach both for the design of new exploration tools with higher expressiveness, and formal assessments and comparisons between different tools.

Keywords

Exploration; Formal Model; Framework; Semi-Structured Data

Resumo

Nunes, Thiago Ribeiro; Schwabe, Daniel (Advisor). **Um Modelo para Exploração de Dados Semiestruturados**. Rio de Janeiro, 2017. 144p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Tarefas de exploração de informação são reconhecidas por possuir características tais como alta complexidade, falta de conhecimento do usuário sobre o domínio da tarefa e incertezas sobre as estratégias de solução. O estado-da-arte em exploração de dados inclui uma variedade de modelos e ferramentas baseadas em diferentes paradigmas de interação, como por exemplo, busca por palavras-chave, busca facetada e orientação-a-conjuntos. Não obstante os muitos avanços das últimas décadas, a falta de uma abordagem formal do processo de exploração, juntamente com a falta de uma adoção mais pragmática do princípio de separação-de-responsabilidades no design dessas ferramentas são a causa de muitas limitações. Dentre as limitações, essa tese aborda a falta de expressividade, caracterizada por restrições na gama de estratégias de solução possíveis, e dificuldades de análise e comparação entre as ferramentas propostas. A partir desta observação, o presente trabalho propõe um modelo formal de ações e processos de exploração, uma nova abordagem para o projeto de ferramentas de exploração e uma ferramenta que generaliza o estado-da-arte em exploração de informação. As avaliações do modelo, realizadas por meio de estudos de caso, análises e comparações o estado-da-arte, corroboram a utilidade da abordagem.

Keywords

Exploração; Modelo Formal; Framework; Dados Semiestrutrados

Summary

List of Images