

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



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

All rights reserved.

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: Prof^ª. Adriana Pereira de Medeiros, Prof. Antonio Luiz Furtado, Prof. Edward Hermann Haeusler, Prof. Hélio Côrtes Vieira Lopes, Prof^ª. 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 Semiestruturados

Summary

1 Introduction	12
1.1. Research Questions	14
1.2. Research Goals and Methodology	15
2 Information Exploration	17
2.3. Summary	28
3 A Framework of Reference	30
3.1. Functional Layer	35
3.2. Interaction/Interface	37
3.3. Data Model	40
3.4. Related Works	42
4 Functional Layer of Exploration Framework	44
4.1. Preliminary notations	44
4.2. The Exploration Process	45
4.3. Data Model	46
4.3.1. Dataset, Items, and Relations	47
4.3.2. Exploration Sets and Exploration Items	50
4.4. A Model of Exploration Operations	52
4.4.1. Notational Convention for Functions	52
4.4.2. Extension-Oriented Operations	53
4.5. Reusing Explorations	69
5 Case Study Evaluations	74
5.1. Case Study 1: Discovering Technological Trends	74
5.1.1. Task Execution	75
5.1.2. Alternative Strategies	77
5.1.3. Generalizing and Reusing Exploration Patterns	79
5.2. Case Study 2: Evaluating a scientific paper	80

5.3. Case Study 3: Summarizing Gene Clusters	84
5.4. Alternative Strategy	88
5.4.1. Generalization and Reuse	89
5.5. Conclusions	89
6 Evaluation of Exploration Tools	91
6.1. Tactical Analysis	91
6.2. Strategic Analysis	99
6.2.1. Notational Conventions	100
6.2.2. Evaluation	102
6.3. Evaluating Business Intelligence and Visualization tools	107
7 Design Issues and Implementation	113
7.1. Functional Layer	113
7.2. Interaction/Interface Design	115
7.2.1. Requirement 1: Manipulation of Exploration Sets and Items	115
7.2.2. Requirement 2: Applying Exploration Operations	122
7.2.3. Requirement 3: Exploration trail management and browsing	129
8 Conclusions and Future Works	133
9 References	136
Attachement A	143

List of Images

Figure 1 - Information Retrieval model (BATES, 1989)	18
Figure 2 - Wilson's Model of Information Seeking (WILSON, T.D., 1999)	23
Figure 3 - Bate's Berrypicking Model (BATES, 1989)	24
Figure 4 - Marchionini's Exploratory Search Model (MARCHIONINI, 2006)	26
Figure 5 - Norman's execution and evaluation gulfs (NORMAN; DRAPER, 1986)	31
Figure 6 - Semantic and articulatory distances for gulf traversals (NORMAN; DRAPER, 1986)	32
Figure 7 - The layers of an exploration environment architecture.	34
Figure 8 - gfacet screenshot	39
Figure 9 - Rhizomer screenshot presenting the filtering relations as facets and the pivoting controls (GARCÍA <i>et al.</i> , 2013)	40
Figure 10 - Nesting of papers by author by publication year.	47
Figure 11 - Relation tree representing publication-author relationships	48
Figure 12 - An exploration set and a path pattern containing the filters for each level.	56
Figure 13 - Representation of a horizontal mapping in the third level	62
Figure 14 - Path example that correlates the Senator <i>Christopher Bond</i> with the state of <i>Missouri</i> in Gov.Track.Us dataset (ARAUJO <i>et al.</i> , 2010)	68
Figure 15 - Abstract path pattern that can be generated by a <i>VerticalMap</i> (ARAUJO <i>et al.</i> , 2010)	68
Figure 16 - Exploration graph for finding research areas in common between researchers affiliated to PUC-Rio and UFRJ with reevaluation ordering.	73
Figure 17 - Patents dataset summarized schema	75
Figure 18 - Open citations summarized schema	81
Figure 19 - Schema of a gene dataset	86
Figure 20 - Tactical Analysis Tree	92
Figure 21 - pivot and refine in gfacet (HEIM; ZIEGLER; LOHMANN, 2008)	93
Figure 22 – gfacet tactical map	94
Figure 23 - (A) Refinement of items by type and relations (movies by	

Title, Year, Director, or Genre). (B) Refined movies and relations presented in a tabular format.	96
Figure 24 - Movies table expanded with theaters and restaurants	96
Figure 25 - Tactical map of SeCo	96
Figure 26 - Tactical map of faceted search tools	98
Figure 27 – Independent refinement executions.	101
Figure 28 – Tableau’s main screen	108
Figure 29 - Tableau's initial table view presenting the table <i>Person</i> (A); A join between the tables <i>Person</i> and <i>OrgPerson</i> .	109
Figure 30 – Filters definition view in Tableau	109
Figure 31 – Tableau’s tactical map	110
Figure 32 - DSL representation of the solution strategy for the paper review case study presented in chapter 5, section 2	114
Figure 33 - The interface of the XPlain environment. (A) keyword search controls; (B) Exploration operations toolbar; (C) Exploration sets area; (D) Exploration trail view.	115
Figure 34 - Visor screenshot (POPOV <i>et al.</i> , 2011)	118
Figure 35 - Tabular view of Liquid Query (BOZZON, ALESSANDRO <i>et al.</i> , 2010)	118
Figure 36 - Rhizomer list view	119
Figure 37 - /facet multilevel list view	119
Figure 38 - Visual representation of an exploration set as a nesting of items and relations.	120
Figure 39 – Xplain’s view for the <i>Refine</i> operation. The user selects relations (A) or relation paths (B) and restriction values for each filter. Filters can be disjunctive or conjunctive according to the selected logical operator.	124
Figure 40 - Example of MoLIC diagram for the task of buying a ticket	126
Figure 41 - (A) Pivot and Refine operations in a single scene; (B) Pivot operation defined in a different interaction scene.	127
Figure 42 - XPlain interaction dialogue for combinations of Pivot and Refine	128
Figure 43 - Linear implementations of Parallax (A) (HUYNH;	

KARGER, 2009) and /facet (B) (HILDEBRAND; OSSENBRUGGEN;
HARDMAN, 2006), and tree view of Sewelis (C) (FERRÉ; HERMANN,
2012) 130

Figure 44 - Graph representation of the functional composition for the task
“finding relevant and not cited papers”. 131