

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO



Carolina Howard Felicíssimo

**Interoperabilidade Semântica na Web:
Uma Estratégia para o Alinhamento Taxonômico de Ontologias**

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Informática da PUC-Rio como parte dos requisitos parciais para obtenção do título de Mestre em Informática.

Orientadores: Julio Cesar Sampaio do Prado Leite
Karin Koogan Breitman

Rio de Janeiro
Agosto de 2004



Carolina Howard Felicíssimo

**Interoperabilidade Semântica na Web:
Uma Estratégia para o Alinhamento Taxonômico
de Ontologias**

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. Julio Cesar Sampaio do Prado Leite

Orientador

Departamento de Informática – PUC-Rio

Profª. Karin Koogan Breitman

Co-orientadora

Departamento de Informática – PUC-Rio

Prof. Carlos José Pereira de Lucena

Departamento de Informática – PUC-Rio

Profª Simone Diniz Junqueira Barbosa

Departamento de Informática – PUC-Rio

Prof. Ricardo Choren Noya

Depto de Engenharia de Computação – IME

Prof. José Eugênio Leal

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de janeiro, 19 de agosto de 2004

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

Carolina Howard Felicíssimo

Graduou-se em Engenharia de Computação na PUC-Rio em 2001. É pesquisadora associada ao Laboratório de Engenharia de Software (LES) da PUC-Rio, atuando na área de Web Semântica da Engenharia de Software.

Ficha Catalográfica

Felicíssimo, Carolina Howard

Interoperabilidade semântica na Web : uma estratégia para o alinhamento taxonômico de ontologias / Carolina Howard Felicíssimo ; orientadores: Julio Cesar Sampaio do Prado Leite, Karin Koogan Breitman. – Rio de Janeiro : PUC, Departamento de Informática, 2004.

180 f. : il. ; 30 cm

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

Inclui referências bibliográficas.

1. Informática – Teses. 2. Web semântica. 3. Interoperabilidade semântica. 4. Ontologias. 5. Interoperabilidade de ontologias. 6. Alinhamento. 7. Agentes de software. 8. Engenharia de software. I. Leite, Julio Cesar Sampaio do Prado. II. Breitman, Karin Koogan. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

A todos aqueles que, de uma forma ou de outra,
ajudaram a fazer este trabalho.

Agradecimentos

A Deus, por todos os caminhos iluminados decisivos na minha vida.

À minha família pelo apoio, carinho, suporte e encorajamento. Em especial, aos meus pais e irmã, à Izalea, ao meu afilhado Jordan e a sua irmã Giulia.

Ao meu orientador, Professor Julio Cesar Sampaio do Prado Leite que, mesmo quando eu ainda estava na graduação, me ofereceu a oportunidade de fazer mestrado sob sua orientação. Sua preocupação com os detalhes ajudou muito para a qualidade deste trabalho.

À minha co-orientadora, Professora Karin Koogan Breitman que me apresentou ao tema de pesquisa da dissertação e teve certeza, desde o início, deste tratar-se de um bom tema. Sua paciência, apoio, atenção e didática contribuíram para que, com o tempo, eu conseguisse caminhar sozinha no desenvolvimento do trabalho com tranquilidade e confiança.

À minha amiga Karin, por me propiciar tantos momentos de descontração com sua família e amigos. Pelo otimismo e confiança na minha capacidade.

Aos meus amigos, em especial, Gustavo Robichez, Miriam Sayão, Lyrene Fernandes e Roberto Martins por todos os ensinamentos profissionais e pessoais, e por estarem sempre presentes em tantos momentos difíceis e bons da minha vida.

Aos Professores Carlos José Pereira de Lucena, Simone Barbosa e Ricardo Choren por participarem da Comissão Examinadora. Ao professor Ulf Bergmann pela ajuda e por todos os recursos disponibilizados, essenciais para a execução do trabalho no período devido. A todos os professores e funcionários do Departamento pela ajuda. Ao Luís Fernando pelo seu suporte.

À CAPES, à FAPERJ e ao Departamento de Informática da PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Resumo

Felicíssimo, Carolina Howard; Leite, Julio Cesar Sampaio do Prado; Breitman, Karin Koogan. Interoperabilidade Semântica na Web: Uma Estratégia para o Alinhamento Taxonômico de Ontologias. Rio de Janeiro, 2004. 180p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Com a evolução da Web atual para a Web Semântica, acredita-se que as informações disponíveis estarão estruturadas de forma a permitir o processamento automático de seu conteúdo por máquinas. Além do processamento individual, deseja-se uma melhor troca de informações entre aplicações Web. Para estes propósitos, são necessários mecanismos que garantam a interoperabilidade semântica, i.e., identificação e compatibilidade de informações. Neste sentido, ontologias são utilizadas como um recurso para disponibilizar um vocabulário estruturado e livre de ambigüidades. Ontologias fornecem um padrão bem definido para a estruturação da informação e promovem um formalismo passível de processamento automático. Neste trabalho, propomos uma estratégia para interoperabilidade de ontologias. O Componente para Alinhamento Taxonômico de Ontologias – CATO, resultado da implementação desta estratégia proposta, alinha automaticamente as taxonomias de ontologias comparadas. O alinhamento realizado é obtido em três etapas executadas seqüencialmente. A primeira etapa compara lexicalmente os conceitos das ontologias entradas e usa um mecanismo de poda estrutural dos conceitos associados como condição de parada. A segunda etapa compara estruturalmente as hierarquias das ontologias identificando as similaridades entre suas sub-árvores comuns. A terceira etapa refina os resultados da etapa anterior classificando os conceitos identificados como *similares* em *bem similares* ou *pouco similares*, de acordo com um percentual de similaridade pré-definido.

Palavras-chave

Web Semântica, Interoperabilidade Semântica, Ontologias, Interoperabilidade de Ontologias, Alinhamento, Agentes de Software, Engenharia de Software.

Abstract

Felicíssimo, Carolina Howard; Leite, Julio Cesar Sampaio do Prado; Breitman, Karin Koogan. Semantic Web Interoperability: One strategy for the Taxonomic Ontology Alignment. Rio de Janeiro, 2004. 180p. Master Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

With the Web evolving towards a Semantic Web, it is believed that the available information will be presented in a meaningful way to allow machines to automatically process its content. Besides the individual processing, a better information exchange among Web applications is desired. For this purpose, mechanisms are called for guarantee the semantic interoperability, that is, the identification and compatibility of information. In this direction, ontologies are used as one resource to make available a structured vocabulary, free of ambiguities. Ontologies provide a well-defined standard to structure the information and to promote formalism for automatic processing. In this work, we propose one strategy for ontology interoperability. The Ontology Taxonomic Alignment Component – CATO, which is the result of the implementation of this proposed strategy, provides an automatic taxonomic ontologies alignment. In this way, the alignment is obtained by a three-step process. The first step is the lexical comparison between the concepts from the entries ontologies. It uses a trimming mechanism of the related associated concepts as a stop condition. The second step is the structural comparison of the ontologies structures used to identify the similarities between common sub-trees. The third step refines the results of the previous step, classifying the *similar* identified concepts as *very similar* or *little similar*, according to a pre-defined similarity measurement.

Keywords

Semantic Web, Semantic Interoperability, Ontologies, Ontologies Interoperability, Alignment, Software Agents, Software Engineering.

Sumário

1 Introdução	14
1.1. A Web Semântica	14
1.2. Interoperabilidade Semântica	16
1.3. Ontologia	17
1.4. Guia do Leitor	20
2 Interoperabilidade de Ontologias	21
2.1. Mecanismos	21
2.2. Alinhamento de Ontologias	24
2.2.1. Compatibilidade de Termos	25
2.2.2. Requisitos	26
2.2.3. Resultados Considerados Satisfatórios	27
2.3. Revisão da Literatura	28
3 A Estratégia	33
3.1. Um Exemplo Simplificado	35
3.2. Primeira Etapa: Comparação Lexical com Uso de Sinônimos e Mecanismo de Poda Estrutural como Condição de Parada	37
3.2.1. Revendo o Exemplo	39
3.3. Segunda Etapa: Comparação Estrutural Usando uma Implementação do Algoritmo <i>TreeDiff</i>	40
3.3.1. Informações de equivalência	43
3.3.2. Grupos de Equivalência	44
3.3.3. Revendo o Exemplo	45
3.4. Terceira Etapa: Uso de Medidas de Similaridades para os Ajustes Finos	46
3.4.1. Revendo o Exemplo	48
3.5. A Implementação	49
3.5.1. Características da linguagem Java	49
3.5.2. A API Jena	50

3.5.3. A linguagem <i>OWL</i>	50
3.5.4. O CATO	51
4 Estudo de Caso	58
4.1. Introdução	58
4.2. Estratégia de Seleção das Ontologias	59
4.3. Primeiro Estudo de Caso	61
4.3.1. Resultado da Primeira Etapa	64
4.3.2. Resultados das Etapas sem Ordenação Alfabética	64
4.3.3. Resultados das Etapas com Ordenação Alfabética	67
4.3.4. Avaliação dos Resultados	70
4.3.5. Problemas Encontrados	71
4.4. Segundo Estudo de Caso	72
4.4.1. Resultado da Primeira Etapa	73
4.4.2. Resultados das Etapas sem Ordenação Alfabética	74
4.4.3. Resultados das Etapas com Ordenação Alfabética	76
4.4.4. Avaliação dos Resultados	78
4.4.5. Problemas Encontrados	79
4.5. Terceiro Estudo de Caso	79
4.5.1. Resultado da Primeira Etapa	81
4.5.2. Resultados das Etapas sem Ordenação Alfabética	82
4.5.3. Resultados das Etapas com Ordenação Alfabética	84
4.5.4. Avaliação dos Resultados	86
4.5.5. Problemas Encontrados	87
5 Conclusões	88
5.1. Contribuições	88
5.1.1. Comparação com outras soluções	89
5.2. Avaliação da Estratégia	90
5.3. Trabalhos Futuros	94
6 Referências Bibliográficas	101

Ontologia de saída do módulo sem ordenação alfabética	108
Ontologia de saída do módulo com ordenação alfabética	108
Primeira Ontologia de entrada	108
Segunda Ontologia de entrada	115
Anexo B – Código em <i>OWL</i> das Ontologias do Segundo Estudo de Caso	137
Primeira Ontologia de entrada	137
Segunda Ontologia de entrada	137
Ontologia de saída do módulo sem ordenação alfabética	137
Ontologia de saída do módulo com ordenação alfabética	137
Anexo C – Código em <i>OWL</i> das Ontologias do Terceiro Estudo de Caso	138
Primeira Ontologia de entrada	138
Segunda Ontologia de entrada	138
Ontologia de saída do módulo sem ordenação alfabética	138
Ontologia de saída do módulo com ordenação alfabética	138
Anexo D – Informações dos Métodos do CATO	139
Anexo E – Código em <i>Java</i> da Implementação do CATO	144
SolCombSinonimos.java e SolCombSinonimosWithOrderNodes.java	144
BDQuery.java	163
DOMComparatorViewWithoutInterface.java	165
LastStepSolCombSinonimos.java	168

Lista de figuras

Figura 1 – Arquitetura definida para a Web Semântica em (Berners-Lee, 2000b)	15
Figura 2 – Combinação de ontologias, adaptado de (Noy, 1999b)	22
Figura 3 – Alinhamento de ontologias, adaptado de (Noy, 1999b)	22
Figura 4 – Mapeamento de ontologias, adaptado de (Noy, 1999b)	23
Figura 5 – Integração de ontologias	23
Figura 6 – Interseção de mercadorias de aplicações complementares	24
Figura 7 – O conjunto de ferramentas <i>PROMPT</i> de (Noy e Musen, 2003)	29
Figura 8 – Estratégia para o alinhamento taxonômico de ontologias	34
Figura 9 – Exemplo de ontologias a serem alinhadas	36
Figura 10 – Informações cadastradas no banco de sinônimos criado	37
Figura 11 – Sinônimos identificados dos conceitos das ontologias analisadas	39
Figura 12 – Informações identificadas na primeira etapa da estratégia	40
Figura 13 – Uso do algoritmo do <i>TreeDiff</i> de (Bergmann, 2002)	42
Figura 14 – Transformação da estrutura de ontologias para a estrutura em <i>XML</i>	43
Figura 15 – Grupos de equivalência identificados no módulo sem ordenação alfabética	45
Figura 16 – Grupos de equivalência identificados no módulo com ordenação alfabética	45
Figura 17 – Valores calculados de similaridade entre os termos	47
Figura 18 – Informações identificadas na terceira etapa da estratégia	49
Figura 19 – Classes dos módulos principais do CATO, com seus métodos	53
Figura 20 – Arquitetura do CATO	54
Figura 21 – Entradas e saídas das etapas de alinhamento do CATO	55
Figura 22 – Representações da ontologia de publicações escolhida	61
Figura 23 – Gerenciamento de Conhecimento no ITM	62
Figura 24 – Ontologias comparadas	63
Figura 25 – Sinônimos cadastrados identificados	64
Figura 26 – Grupos de equivalência identificados no módulo sem ordenação	66
Figura 27 – Percentuais de similaridade calculados	66

Figura 28 – Informação adicionada, resultado do alinhamento com o CATO	67
Figura 29 – Grupos de equivalência identificados no módulo com ordenação alfabética	68
Figura 30 – Percentuais de similaridade calculados	69
Figura 31 – Informação adicionada, resultado do alinhamento com o CATO	70
Figura 32 – Ontologias comparadas	72
Figura 33 – Sinônimos cadastrados identificados	74
Figura 34 – Grupos de equivalência identificados no módulo sem ordenação	74
Figura 35 – Percentuais de similaridade calculados	75
Figura 36 – Informação adicionada, resultado do alinhamento com o CATO	76
Figura 37 – Grupos de equivalência identificados no módulo com ordenação alfabética	77
Figura 38 – Percentuais de similaridade calculados	77
Figura 39 – Informação adicionada, resultado do alinhamento com o CATO	78
Figura 40 – Ontologias comparadas	80
Figura 41 – Sinônimos cadastrados identificados	81
Figura 42 – Informação adicionada, resultado do alinhamento com o uso de sinônimos na primeira etapa do CATO	82
Figura 43 – Estruturas hierárquicas das ontologias comparadas	83
Figura 44 – Grupos de equivalência identificados no módulo com ordenação alfabética	85
Figura 45 – Percentuais de similaridade calculados	85
Figura 46 – Informação adicionada, resultado do alinhamento com o CATO	86
Figura 47 – Sinônimos do termo carro, ordenados por estimativa de frequência	96
Figura 48 – Hiperonímia de carro significando veículo automotor (carro é um tipo de ...)	96
Figura 49 – Hiponímia de carro significando veículo automotor (... é um tipo de carro)	97
Figura 50 – Holonímia de carro (carro é parte de ...)	97
Figura 51 – Meronímia de carro significando veículo automotor (... é parte de carro)	97
Figura 52 – Termos do domínio de carro significando veículo automotor	97
Figura 53 – Representação em árvore de relacionamentos de composição	98

Lista de Abreviaturas

API – *Application Programming Interface*
ATLAS – *Agent Transaction Language for Advertising Services*
CATO – *Componente para Alinhamento Taxonômico de Ontologias*
CMU – *Carnegie Mellon University*
DAML – *Darpa Agent Markup Language*
OIL – *Ontology Inference Layer*
DAML+OIL – *Darpa Agent Markup Language + Ontology Inference Layer*
DOM – *Document Object Model*
HTML – *HyperText Markup Language*
IEEE – *Institute of Electrical and Electronics Engineers*
ITM – *Intelligent Topic Manager*
LAL – *Léxico Ampliado da Linguagem*
OWL – *OWL Web Ontology Language*
PUC-RIO – *Pontifícia Universidade Católica do Rio de Janeiro*
RDF – *Resource Description Framework*
RDF Schema – *RDF Vocabulary Description Language*
SUMO – *Suggested Upper Merged Ontology*
Udl – *Universo de Informação*
URI – *Uniform Resource Identifier*
URL – *Uniform Resource Locator*
XML – *Extensible Markup Language*
W3C – *World Wide Web Consortium*
Web – *World Wide Web*