

Universidade Federal de Pernambuco  
Centro de Informática

Graduação em Engenharia da Computação

**Análise comparativa de técnicas de  
seleção de protótipos**

Dayvid Victor Rodrigues de Oliveira

Trabalho de Graduação

Recife  
22 de novembro de 2011

Universidade Federal de Pernambuco  
Centro de Informática

Dayvid Victor Rodrigues de Oliveira

## **Análise comparativa de técnicas de seleção de protótipos**

*Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.*

Orientador: *Prof. Dr. George Darmiton*

Recife  
22 de novembro de 2011

*Eu dedico este trabalho a João Rodrigues de Silva, meu  
avô.*

# Agradecimentos

Senhor, obrigado por ter me dado a fora e a confiana para completar esta jornada,  
obrigado por ter me guiado atravs de todos os obstculos no meu caminho,  
e por me manter  
obrigado pela tua proteo em todo caminho,  
obrigado pelos amigos que eu fiz,  
que o Senhor cuide deles, assim como cuidou de mim,  
obrigado por me permitir (...)

*What can I give back to God, for the blessings You pour out on me?*

—BONO (Boston, 2001)

# Resumo

RESUMO

**Palavras-chave:** PORTUGUES

# Abstract

ABSTRACT

**Keywords:** INGLES

# Sumário

0.1	Web Semântica	1
0.1.1	Aplicações	2
0.1.1.1	Gerenciamento de Conhecimento	2
0.1.1.2	Comércio eletrônico <i>Business to Consumer</i> (B2C)	2
0.1.1.3	Comércio eletrônico <i>Business to Business</i> (B2B) e agentes pessoais	2
0.1.2	Tecnologias	3
0.1.2.1	Metadados Explícitos	3
0.1.2.2	Ontologias	4
0.1.2.3	Lógica	4
0.1.2.4	Agentes	5
0.2	Organização da Monografia	5
<b>1</b>	<b>Web Ontology Language: OWL</b>	<b>6</b>
1.1	Definition 1 (First-order logic syntax).	6



# Lista de Figuras

1	Código HTML de uma página da americanas.com	3
2	Exemplo de código com semântica para um produto	4

# **Lista de Tabelas**

A World Wide Web é uma das tecnologias mais revolucionárias que o homem já inventou. Ela mudou em escala global a forma com que pessoas e empresas trocam informações, contribuindo para que o conhecimento se tornasse mais universal e que limites físicos e lingüísticos fossem cada vez mais minimizados.

A web como conhecemos hoje nasceu de uma proposta feita por Tim Berners-Lee à empresa CERN em 1989 [?]. O problema enfrentado pela empresa na época era a perda de informações internas por falta de documentação ou pela saída de algum funcionário. A solução proposta por Berners-Lee foi fazer uma rede de documentos interligados por hyperlinks em que cada setor da empresa poderia adicionar novos documentos.

A estrutura básica que Berners-Lee montou a 22 anos evoluiu a passos largos em relação à escalabilidade e padronização de protocolos e linguagens, tendo hoje cerca de 2 bilhões de usuários, mais de 30% da população do planeta.

Apesar do avanço das infra-estruturas e serviços para a Web, ainda há muito o que evoluir. Uma das propostas de mudanças é prover uma maior expressividade da linguagem que descreve os documentos na Web [?]. Hoje, esses documentos não possuem um significado que possa ser extraído de forma concisa, apresentam ambigüidade, misturam os dados com elementos visuais e muitas vezes não podem ser indexados por engenhos de busca.

## 0.1 Web Semântica

*"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web, the content, links, and transactions between people and computers. A **Semantic Web** which should make this possible has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The **intelligent agents** people have touted for ages will finally materialize."*Tim Berners-Lee

Tradução literal: *"Eu tenho um sonho para a Web [em que os computadores] tornam-se capazes de analisar todos os dados na Web, o conteúdo, links, e as transações entre pessoas e computadores. A **Web Semântica** que deve tornar isso possível ainda está para surgir; mas quando isso acontecer, os mecanismos dia-a-dia da burocracia do comércio e nossas vidas diárias serão tratados por máquinas falando com máquinas. Os **agentes inteligentes** que as pessoas têm falado por anos vão finalmente se concretizar."*Tim Berners-Lee

A Web Semântica citada no texto de Berners-Lee acima é uma iniciativa de pesquisadores da área de inteligência artificial e lingüística computacional que estudam como adequar a Web de hoje a uma infra-estrutura que a tornará mais acessível às máquinas. Essa nova roupagem que os pesquisadores querem dar à Web permitirá que serviços mais sofisticados possam ser construídos, como os que serão descritos a seguir.

### 0.1.1 Aplicações

#### 0.1.1.1 Gerenciamento de Conhecimento

Gerenciamento de conhecimento está relacionado à aquisição, acesso e manutenção de conhecimento dentro de uma empresa ou organização. Essa atividade se tornou e está se estabelecendo como uma necessidade básica em grandes empresas visto que o conhecimento que é gerado internamente agrega valor, pode se tornar um diferencial competitivo e também pode aumentar a produtividade de seus colaboradores. Com o uso de tecnologias criadas para a Web Semântica, soluções para G.C. podem melhorar em vários aspectos, entre eles:

- Organização do conhecimento existente a partir de seu significado;
- Geração de novas informações de forma automática;
- Checagem de inconsistências semânticas em documentos;
- Substituição de consultas baseadas em palavras-chave por perguntas em linguagem natural;

#### 0.1.1.2 Comércio eletrônico *Business to Consumer* (B2C)

O comércio eletrônico entre vendedores e consumidores é um dos modelos de negócio na Internet que melhor se estabeleceu, sites como amazon <sup>1</sup>, americanas <sup>2</sup> e mercado livre <sup>3</sup> possuem público fiel e que os visitam por vários objetivos. É muito comum para a geração que cresceu imersa na Web entrar em sites de compra como esses a procura do melhor preço antes de decidir fazer uma compra. Muitas vezes o produto não é adquirido em uma loja virtual, mas a pesquisa inicial de preços é que muitas vezes determina a escolha do produto. Observando esse comportamento, sites como o buscapé <sup>4</sup> fazem o trabalho de indicar qual é a loja que está com o melhor preço.

A Web Semântica pode ajudar nesse cenário provendo interfaces de consulta mais completas aos sites que fazem comparação de preços, porém, com muito mais detalhes técnicos sobre o produto. Supondo que cada produto tem, por exemplo, uma ontologia que o descreve em detalhes (provida pelo fabricante ou por sites de review de produtos), o consumidor poderá fazer comparações muito mais detalhadas, ajudando-o a encontrar o produto que vai suprir a sua necessidade.

#### 0.1.1.3 Comércio eletrônico *Business to Business* (B2B) e agentes pessoais

A maioria das pessoas que compram serviços ou produtos na Web só conhecem o comércio eletrônico do tipo B2C, mas existem tecnologias para comércio do tipo B2B, Business to Business, agentes computacionais de empresas que se comunicam para fechar acordos e otimizar o ciclo de negócios que muitas vezes já podem ser previstos e modelados.

---

<sup>1</sup> site: amazon.com

<sup>2</sup> site: americanas.com

<sup>3</sup> site: mercadolivre.com.br

<sup>4</sup> site: buscape.com.br



```
../imagens/americanas.png
```

**Figura 1** Código HTML de uma página da americanas.com

Com a popularização da Web Semântica e a introdução de agentes pessoais e agentes que representam negócios, eles poderão se comunicar de forma mais natural e aplicações para otimizar tarefas do dia-a-dia que poderão ser produzidos. Por exemplo, um médico que possua um agente pessoal que negocie a sua agenda com os agentes pessoais de seus clientes pode ser utilizado para remarcar seus atendimentos em caso de uma viagem ou imprevisto, agindo como uma secretária virtual.

### 0.1.2 Tecnologias

Aplicações como as citadas acima já existem, mas o trabalho de engenharia para conseguir bons resultados é alto devido às tecnologias que são adotadas hoje. Vamos usar o case do *site* de comparação de preços BuscaPé na próxima seção da monografia.

#### 0.1.2.1 Metadados Explícitos

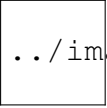
A primeira tarefa de engenharia de um agente coletor de preço, como os do buscaPé, é fazê-lo visitar vários sites de compras todo dia à procura de modificações nas listas de produtos para saber quais estão disponíveis naquela loja. É feito então o *parsing* do HTML de cada site de compras à procura das informações de preço, descrição, avaliação e detalhes de cada produto. A limitação dessa abordagem é que sempre que um dos sites de compra mudar o *layout* (estrutura do HTML), um novo script de *parsing* deverá ser escrito. A grande demanda técnica de uma aplicação como essa é a escrita de agentes muito especializados para atingir bons resultados.

Na figura 1.1 está parte do código em HTML de uma página de produtos da americanas.com. As informações do produto estão cercados apenas de código para a renderização dessa página pelo navegador. Ou seja, a única preocupação dos engenheiros da americanas.com foi a leitura por humanos da informação do produto. Uma aplicação que deseje usar as informações dos produtos da loja vai ter que fazer um agente especializado no *parsing* desse código.

Motores de busca que também se baseiam em *parsing* de páginas para extrair informações da Web dificilmente saberão, por exemplo, qual é o preço de um produto nesse site da americanas.com, já que há várias informações de preço na página e o *parsing* que é feito não é otimizado para sites específicos.

A consequência para o usuário final é que ele terá que usar um site específico como o buscapé ou terá que fazer buscas a um engenho de busca por palavras-chave para achar os sites de compra que possuam um produto e em uma segunda etapa, fazer a análise de preços manualmente.

A abordagem da Web Semântica para resolver problemas como esse não é fazer agentes especializados (como os do buscapé), e sim, anotar metadados semânticos dos documentos disponíveis na Web. O exemplo dado anteriormente seria escrito na figura 1.2.



```
../imagens/americanas-xml.png
```

**Figura 2** Exemplo de código com semântica para um produto

### 0.1.2.2 Ontologias

O termo ontologia vem da filosofia, nesse contexto, é um ramo da filosofia que se dedica a estudar a natureza da existência, concentra-se em identificar e descrever o que existe no universo. Em computação, uma ontologia é um artefato para descrever um domínio. Consiste em uma lista finita de termos e relações entre eles. Os termos denotam conceitos importantes de um domínio [?].

Grande parte dos trabalhos referentes à Web Semântica estão ligados a ontologias, inclusive este. As linguagem de descrição de ontologias mais importantes para a Web são:

- XML: usado para dirigir a sintaxe de documentos estruturados. Não impõe restrições semânticas no conteúdo do documento;
- XML Schema: linguagem para impor restrições na estrutura dos documentos XML;
- RDF: modelo de dados para recursos (objetos) e relações entre eles. As restrições semânticas são fixas e podem ser representados a partir da sintaxe do XML;
- RDF Schema: descreve as propriedades e classes dos objetos RDF;
- OWL: linguagem rica para modelagem de classes, propriedades, relações entre classes (e.g. disjunção), restrições de cardinalidade, características de propriedades (e.g. simetria). Mais detalhes sobre a OWL serão dados no capítulo 2 dessa monografia.

### 0.1.2.3 Lógica

Lógica é a disciplina que estuda os princípios do raciocínio. Ela provê linguagens formais para expressar conhecimento, a semântica formal para a interpretação de sentenças sem precisar realizar operações sobre a base de conhecimento e a transformação de conhecimento implícito em conhecimento explícito, através de deduções a partir da base de conhecimento [?].

Lógica é mais geral que ontologias, ela pode ser usada por agentes inteligentes para tomada de decisões e escolha de ações. Por exemplo, um agente de B2C pode dar um desconto a um cliente baseado na seguinte regra:

$$\forall x \forall y, cliente(x) \wedge produto(y) \wedge clienteFiel(x) \rightarrow desconto(x, y, 5\%)$$

Onde *cliente(x)* indica que x é um cliente/consumidor, *produto(y)* indica que y é um produto de uma loja, *clienteFiel(x)* indica que x é um cliente fiel da loja e *desconto(x, y, 5%)* indica que o cliente x terá um desconto de 5% no produto y.

#### 0.1.2.4 Agentes

Um agente é tudo o que pode ser considerado capaz de perceber seu ambiente por meio de sensores e de agir sobre esse ambiente por meio de atuadores [?]. Agentes lógicos são aqueles que executam ações através de uma base de conhecimento e possuem um requisito fundamental, quando ele formula uma pergunta para a base de conhecimento, a resposta deve seguir o que já foi informado anteriormente.

Agentes para a Web Semântica utilizam as três tecnologias que já foram descritas:

- Metadados serão usados para identificar e extrair informações da Web;
- Ontologias serão usadas para dar assistência às consultas realizadas à Web, interpretar informações recuperadas e para comunicação com outros agentes;
- Lógica será usada para processar informações recuperadas, chegar a conclusões e tomar decisões;

## 0.2 Organização da Monografia

Esta monografia está dividida em cinco capítulos. No Capítulo 1, é apresentada uma visão geral sobre a Web Semântica, exemplificando com aplicações e citando as tecnologias que estão sendo usadas. No Capítulo 2, são apresentados conceitos referentes a Lógica de Descrição e a sua ligação com a linguagem de descrição de ontologias OWL. No Capítulo 3 são descritos os algoritmos para normalização de ontologias para a Forma Normal Positiva. No Capítulo 4 o LeanCop é apresentado e é mostrada a validação do trabalho realizado. O Capítulo 5 apresenta as considerações finais sobre o trabalho, bem como propostas de trabalhos futuros.

# Web Ontology Language: OWL

In order to describe the connection method as a formal inference system (in section 4), and the positive matricial normal form used in it, we will briefly describe the notation we use for first order logic, before examining the method. We are presuming readers to be acquainted to first order logic.

## 1.1 Definition 1 (First-order logic syntax).

The alphabets given by table 1 compose the FOL syntax notation. ? Table 1. First order logic syntax notation.



## Referências Bibliográficas

- [CPZ11] Ruiqin Chang, Zheng Pei, and Chao Zhang. A modified editing k-nearest neighbor rule. *JCP*, 6(7):1493–1500, 2011.
- [dSPC08] Cristiano de Santana Pereira and George D. C. Cavalcanti. Prototype selection: Combining self-generating prototypes and gaussian mixtures for pattern classification. In *IJCNN*, pages 3505–3510. IEEE, 2008.
- [EJJ04] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004.
- [FHA07] Hatem A. Fayed, Sherif Hashem, and Amir F. Atiya. Self-generating prototypes for pattern classification. *Pattern Recognition*, 40(5):1498–1509, 2007.
- [Har68] P. E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.
- [HKN07] Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In Zoubin Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 935–942. ACM, 2007.
- [Koh86] Teuvo Kohonen. Learning vector quantization for pattern recognition. Report TKK-F-A601, Laboratory of Computer and Information Science, Department of Technical Physics, Helsinki University of Technology, Helsinki, Finland, 1986.
- [Koh88] Teuvo Kohonen. Learning vector quantization. *Neural Networks*, 1, Supplement 1:3–16, 1988.
- [PI69] E. A. Patrick and F. P. Fischer II. A generalization of the k-nearest neighbor rule. In *IJCAI*, pages 63–64, 1969.
- [TC67] P.E. Hart T.M. Cover. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21 – 27, 1967.
- [Tom76] I. Tomek. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(2):679–772, 1976.
- [WP01] G. Weiss and F. Provost. The effect of class distribution on classifier learning, 2001.