Universidade de São Paulo
Instituto de Matemática e Estatística
MAC 5778 - Sistemas Baseados em Conhecimento

PR-OWL:
Probabilistic Web Ontology Language

Autor:

Walter Perez Urcia

São Paulo

Novembro 2015

Resumo

Neste trabalho o objetivo é explicar a linguagem PR-OWL desde seu origem até o tempo atual. O artigo começa descrevendo a aparição das ontologias na área e os problemas que tem e que levaram a criar um novo tipo de ontologias que usam probabilidades. Após isso, serão explicadas as diferentes representações que foram usadas para tentar solucionar aqueles problemas e finalmente mostrar a linguagem PR-OWL e suas aplicações.

1 Introdução

O conceito da web semântica é uma ideia de Tim Berners-Lee: "Web semântica é uma extensão da web comum em que a informação tem um melhor definido significado para os computadores e a cooperação das pessoas" [7]. Em outras palavras, mudar a interconexão da web comum a um gigantesco banco de dados relacionado como é mostrado na Figura 1 passando de uma abordagem orientada a arquivos (izquerda) a uma abordagem orientada a informação (direita).

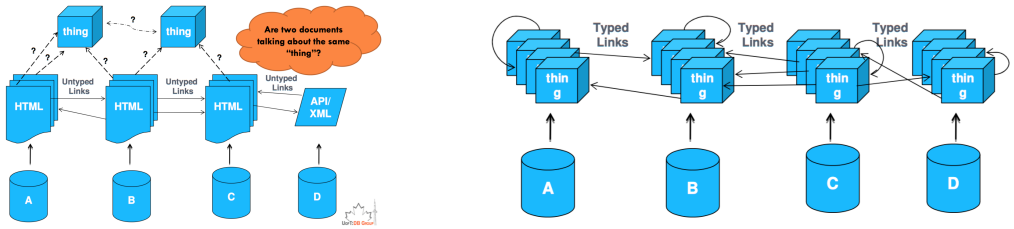

Figura 1: Ideia da web semântica

O problema com aquela ideia de web semântica é que na web atual existem as seguintes situações:

- Amplidão: Existem milhões de páginas
- Não definição única de termos: Existem termos como alto e jovem que são subjetivos
- Incerteza: Conceitos que tem incerteza, por exemplo sintomas de uma doença podem ser de alguma outra com outra probabilidade
- Inconsistência: Existem contradições lógicas
- Engano: A informação obtida não é sempre confiável

Ao longo deste trabalho, serão explicada a principal representação usada para a web semântica: ontologias, seus limitações e como aquele conceito foi extendido a ontologias probabilísticas para lidar com os desafios mencionados anteriormente. Finalmente, serão explicados alguns estudos feitos com aquele último conceito.

2 Ontologias

Uma ontologia é uma representação formal de conhecimento sobre um domínio e está baseada na lógica de descrição. Inclui as seguintes componentes:

- Tipos de entidades (e.g. Pessoa, Companhia, etc)
- Propiedades de entidades (e.g. nome, sobrenome, etc)
- Relações entre entidades (e.g. paiDe, irmãoDe, etc)
- Eventos que acontecem com entidades (e.g. escolher a melhor opção, etc)

Além disso, por ser baseada em lógica, existe a possibilidade de fazer inferências sobre os fatos estabelecidos nela. Também pode ser usada para ajudar os motores de busca responder a perguntas mais complexas.

Exemplo 1: Uso de OWL

```
1  <owl:Class rdf:ID="WineDescriptor" />
2
3  <owl:Class rdf:ID="WineColor">
4      <rdfs:subClassOf rdf:resource="#WineDescriptor" />
5       ...
6  </owl:Class>
7
8  <owl:ObjectProperty rdf:ID="hasWineDescriptor">
9      <rdfs:domain rdf:resource="#Wine" />
10     <rdfs:range  rdf:resource="#WineDescriptor" />
11 </owl:ObjectProperty>
12
13 <owl:ObjectProperty rdf:ID="hasColor">
14     <rdfs:subPropertyOf rdf:resource="#hasWineDescriptor" />
15     <rdfs:range rdf:resource="#WineColor" />
16       ...
17 </owl:ObjectProperty>
```

A linguagem padrão para representar ontologias é chamado OWL (Web Ontology Language) e foi estabelecido por o World Wide Web Consortium (W3C) no ano 2004. Seu uso é mostrado no Exemplo 1 onde estão instanciadas classes, subclasses e propriedades. O problema é que manter aquela estrutura manualmente é complicado e por isso foram aparecendo ferramentas gráficas que a geram automaticamente. A ferramenta mais usada atualmente para construir ontologias é Protegé (Figura 2).

Mas o principal problema de usar as ontologias comuns e OWL como meios para obter a web semântica é que a web atual tem incerteza em muitos conceitos. Esta poderia estar não só em tipos de entidades (e.g. Python é uma linguagem de programação ou um animal), se não também em propriedades ou relações que tem diferentes comportamentos em diferentes domínios (e.g. limpo pode significar que não tem registros policiais, ou que não está sujo). Tendo isto em consideração, existe uma necessidade de encontrar uma representação nova que use tanto lógica como teoria de probabilidades para representar o conhecimento e também de extender a linguagem OWL para ter suporte de probabilidades.

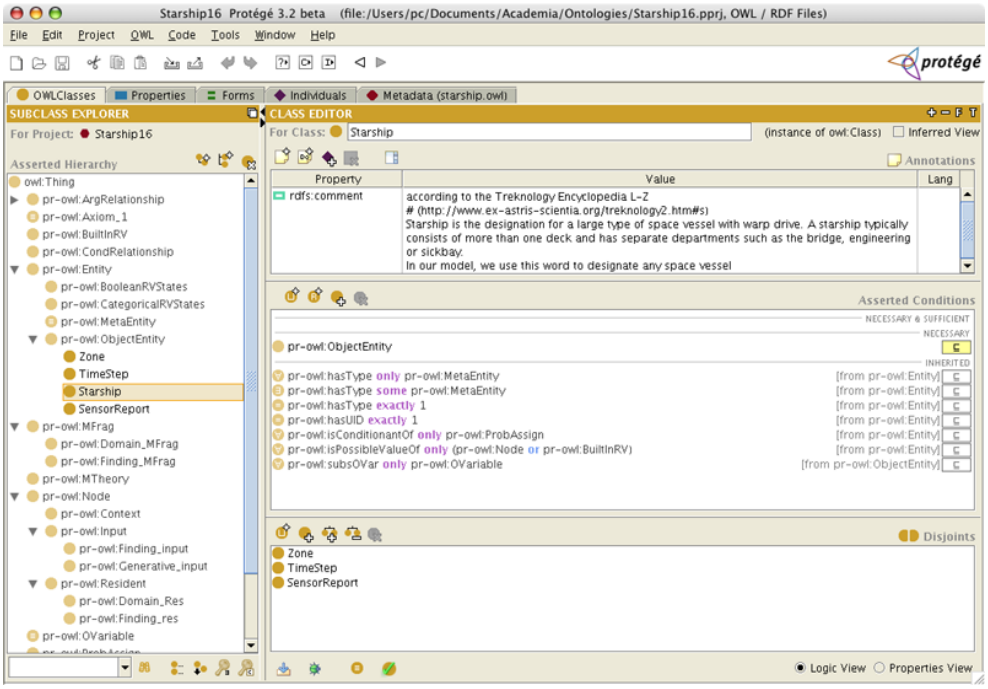

Figura 2: Protegé

3 Ontologias Probabilísticas

Para lidar com a incerteza da web atual foi desenvolvido o conceito de Ontologias Probabilísticas que além de ter as mesmas componentes que as mencionadas na seção 2, são adicionadas algumas características:

- Regularidades estatísticas particulares do domínio
- Conhecimento ambíguo, incompleto e não confiável relacionado às entidades
- Incerteza sobre todas as formas anteriores de conhecimento (e.g. propriedades, classes, etc.)

Durante alguns anos, foram consideradas algumas representações existentes na época para tentar representar ontologias probabilísticas como será visto nas seguintes subseções, mas cada uma delas tinha problemas para lidar tanto com lógica como incerteza [3].

3.1 Redes Bayesianas

Uma rede Bayesiana é um modelo gráfico sobre variáveis (aleatórias) conformada por um grafo direcionado acíclico e uma distribuição de probabilidades conjunta sobre todas as variáveis, estabelecendo independenças condicionais entre elas. Apesar de usar probabilidades e poder lidar com incerteza, sua estrutura é limitada por uma representação atributo-valor. Isto não permite que sejam usadas sentenças gerais da lógica (e.g. $paiDe(X)$).

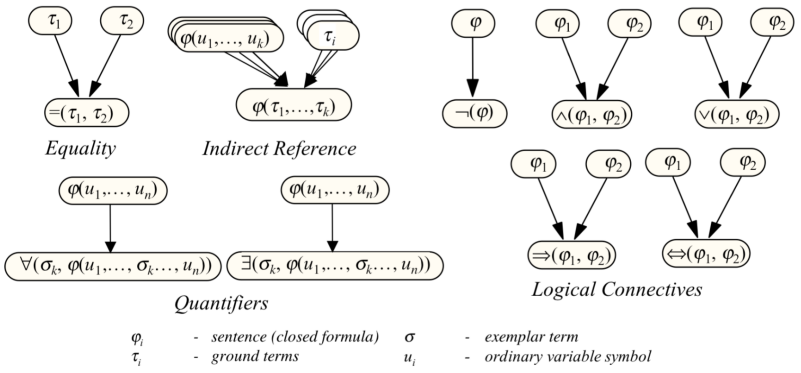

Figura 3: Logical MFrags

3.2 Modelos Ocultos de Markov

Um modelo oculto de Markov é um caso particular de redes Bayesianas dinâmicas. Este tipo de redes adicionam uma variável de tempo a cada um dos valores nela. Apesar que Isto permite que podam ser usadas sentenças recursivas como na lógica, como no caso anterior, esta estrutura também está limitada pela representação atributo-valor.

3.3 Modelos relacionados probabilísticos

Estes modelos extendem as redes Bayesianas para trabalhar com tipos de entidades, que é um dos componentes das ontologias e que as estruturas anteriores não conseguiam fazer diretamente. Ainda assim, o problema é que não conseguem ser usadas para representar sentenças com quantificadores (e.g. $\forall paiDe(X, Y) \rightarrow filhoDe(Y, X)$).

3.4 Multi-Entity Bayesian Networks

Finalmente, no ano 2008, Laskey conseguiu desenvolver a primeira representação que combina lógica com incerteza de forma satisfatória. Aquela representação está baseada em um tipo de rede Bayesiana chamado Multi-Entity Bayesian Network (MEBN) onde cada sentença em lógica de primeira ordem pode ser representada como um fragmento ou MFrag como se mostra na Figura 3. Portanto, permite variáveis não instanciadas, quantificadores e recursão. Cada um desses fragmentos tem distribuições de probabilidades locais e o conjunto de MFrags construi toda a base de conhecimento chamada MTheory [4]. Na seguinte seção será usada esta estrutura para desenvolver a linguagem padrão PR-OWL para representar ontologias probabilísticas.

Um MFrag pode ser definido com 5 componentes:

- C: Conjunto finito de valores de contexto

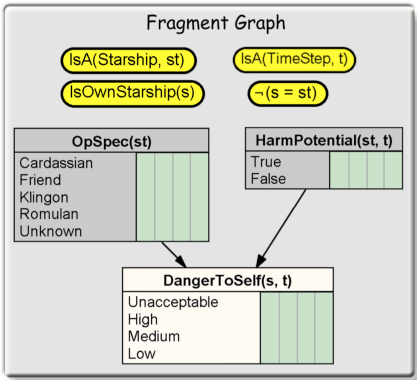

Figura 4: DangerToSelf MFrag

- I: Conjunto finito de termos de variáveis de entrada
- R: Conjunto finito de termos de variáveis residentes
- G: Um grafo direcionado acíclico
- D: Distribuições de probabilidades locais para cada variável no conjunto R

Por exemplo, na figura 4, os valores de contexto estão em cor amarelo, em cinza as variáveis de entrada e em branco as variáveis residentes. As tabelas em cada nó do grafo representam suas distribuições de probabilidades locais.

4 PR-OWL

Usando MEBNs foi desenvolvida a extensão da linguagem OWL com incerteza, chamada PR-OWL (Probabilistic Web Ontology Language) que conseguia adicionar probabilidades a conjuntos de variáveis que dependem de outras. Dito de outra forma, para cada sentença lógica representada em um MEBN e suas instâncias possíveis, podiam ser adicionadas probabilidades [1].

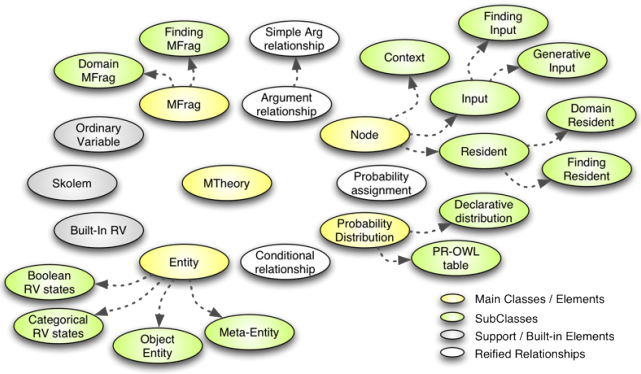

Figura 5: PR-OWL 1.0

Em sua primeira versão chamada PR-OWL 1.0, foi criada uma super ontologia mostrada na figura 5 que fazia uma simulação da estrutura de um MEBN. Aquela super ontologia podia ser usada em Protegé para construir ontologias probabilísticas. Mas como se ve na figura 2, usá-la nessa ferramenta era complexo pela sua estrutura.

Para solucionar aquele problema, no ano 2010, apareceu PR-OWL 2.0, que podia ser usada só na ferramenta UnBBayes [6] (Figura 6, desenvolvida exclusivamente para construir ontologias probabilísticas do mesmo jeito que Protegé fazia com as ontologias comuns. Apesar disso, UnBBayes permite exportar arquivos .owl que são compatíveis para ser usados em Protegé, mas que tem a estrutura da super ontologia anteriormente mencionada.

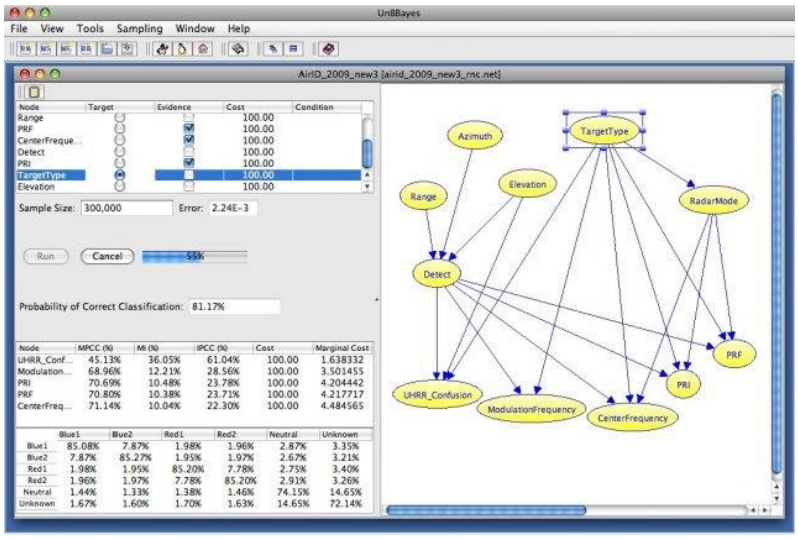

Figura 6: Interface gráfica de UnBBayes

5 Estudos e aplicações

Por último, com PR-OWL já definido formalmente e obtida uma ferramenta para construir ontologias probabilísticas foram desenvolvidos alguns estudos em situações do mundo real como o modelamento de uma ontologia marítima [5] e outra para o reconhecimento de fraudes em Brasil [2].

5.1 Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil

Para lidar com as demandas dos cidadãos de transparência e prevenção de corrupção, a Controladoria-Geral da União organizou campanhas para educar a gente nesses temas e fiz inspeções a múltiplas instituições. Mas apesar de ter toda essa informação coletada de vários lugares (municipalidades, a Polícia Federal, etc), não tinham uma

forma eficiente de unir todas elas. Além disso, a deteção de fraudes é feito manualmente por um auditor e o número de casos que pode analizar durante um tempo é limitado. O principal problema que eles tinham era a incerteza em muitos dos termos de todas as instituições. Este trabalho explica o proceso da construção de uma ontologia probabilística usando PR-OWL 2.0 para detectar fraudes em aquisições em Brasil de forma automática.

5.2 PR-OWL 2 Case Study: A Maritime Domain Probabilistic Ontology

Este trabalho mostra o processo necessário para mudar uma ontologia comum a uma ontologia probabilística usando como exemplo um domínio marítimo (Figura 7) para responder os seguintes queries:

- O navio tem um terrorista?
- O navio tem uma rota não comum?
- O navio parece ter uma atitude evasiva a ataques?
- Qual é o tipo de aquele navio?
- Que bandeira pode ter aquele navio?

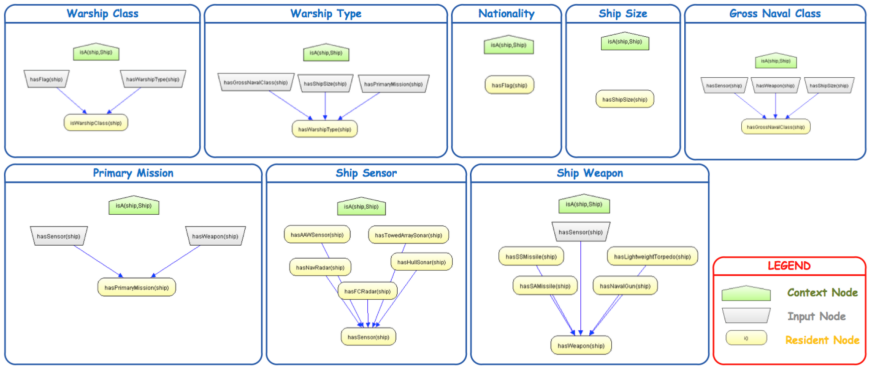

Figura 7: Ontologia Probabilística de domínio marítimo

6 Conclusões

As ontologias probabilísticas ajudam a lidar com a incerteza, que é um dos principais desafios para conseguir a web semântica e que as ontologias comuns não conseguiam abordar. Além disso, a linguagem PR-OWL 1 está baseada na criação de uma super ontologia para ser usada nas ferramentas que usam OWL, o que adiciona mais complexidade para construir uma ontologia probabilística. Por outro lado, PR-OWL 2 é totalmente compatível com a linguagem OWL e as ferramentas que usam ela, mas

ajuda a ter ferramentas desenvolvidas exclusivamente para ontologias probabilísticas como UnBBayes e isso ajudou os pesquisadores da área a realizar mais estudos nos últimos anos.

Referências

[1] PR-OWL Official Site, howpublished = `http://www.pr-owl.org`, note = Accessed: 2015-09-15.

[2] RommelN. Carvalho, Shou Matsumoto, KathrynB. Laskey, PauloC.G. Costa, Marcelo Ladeira, and LaécioL. Santos. Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil. In *Uncertainty Reasoning for the Semantic Web II*, volume 7123 of *Lecture Notes in Computer Science*, pages 19–40. Springer Berlin Heidelberg, 2013.

[3] Paulo C. G. da Costa and Kathryn Blackmond Laskey. Multi-entity bayesian networks without multi-tears, 2010.

[4] Kathryn B. Laskey. MEBN: A Language for First-ORder Bayesian Knowledge Bases. *Artificial Intelligence*, 172, 2008.

[5] Kathryn Blackmond Laskey, Richard Haberlin, Paulo Costa, and Rommel Novaes Carvalho. PR-OWL 2 Case Study: A Maritime Domain Probabilistic Ontology, 2011.

[6] Show Matsumoto Marcelo Ladeira Paulo Costa Rommel Carvalho, Laecio Santos. UnBBayes-MEBN: Comments on Implementing a Probabilistic Ontology Tool. *IADIS Applied Computing 2008 conference*, 2008.

[7] James Hendler Tim Berners-Lee and Ora Lasilla. The Semantic Web. *Scientific American*, 2001.