Desenvolvimento de uma Ontologia

José Pereira PG 27748, Universidade do Minho 16/07/2015

Introdução

Uma ontologia é uma representação formal de conhecimento. Com ela é possível descrever universos complexos de uma forma relacional bastante simples. Neste documento, é descrita uma destas representações no contexto da computação. O domínio estudado é referente à plataforma concreta de comércio de vídeo jogos *Steam*.

0.1 Steam

Atualmente a *Steam* conta com aproximadamente 65 milhões de utilizadores ativos e tem médias de acesso diário de 6,6 milhões de contas em simultâneo. A plataforma também possuí um sistema de amigos, que permite ao utilizador manter uma rede de contactos, com os quais pode jogar, trocar itens, comunicar, etc.
Para esta tarefa, apenas foi considerada a componente de distribuição de jogos desta plataforma, não incluindo qualquer informação relativa a utilizadores.

1 Objectivos

Para construir o produto requerido, foi necessário atingir alguns objetivos mais gerais e que foram levantados à partida:

- Adquirir informação para povoamento da ontologia;
- Esquematizar o domínio de conhecimento de acordo com os dados aglomerados;
- Explorar o potencial de aprendizagem da representação.

2 Abordagem

Nesta secção é descrito o processo de construção da ontologia em detalhe. No percurso do desenvolvimento foram utilizadas as ferramentas:

- *Scrapy* - Ferramenta que implementa uma solução de *web crawling*[1] em *Python*;
- *XSLT* - Linguagem de marcação que define a apresentação de documentos *XML*;
- *Protégé* - Plataforma *open-source* para desenvolvimento e manipulação de modelos de domínio.

2.1 Captura de informação relevante

O estudo de casos concretos e reais enriquece claramente o processo, daí a escolha de um domínio fiel. Para isto, e depois de alguma análise foi estipulado que a informação trabalhada seria obtida através das páginas HTML públicas destinadas ao comércio de jogos. Para capturar tal informação, recorreu-se a uma ferramenta baseada em *web crawling*.

Listing 1: Lista de *URLs* iniciais ao qual será aplicado um parsing e domínio sob o qual é permitido navegar

```
start_urls = ["http://store.steampowered.com/tag/pt/Desporto#p=0&tab=NewReleases",
                  "http://store.steampowered.com/genre/Free%20to%20Play/",
                  "http://store.steampowered.com/genre/Early%20Access/",
                  "http://store.steampowered.com/tag/pt/Aventura/#p=0&tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/Casual/#p=0&tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/Corridas/#p=0&tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/Estrat%C3%A9gia/#p=0&tab=↩
                      NewReleases",
                  "http://store.steampowered.com/tag/pt/Indie/#p=0&tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/Multijogador%20em%20Massa/#p=0&↩
                      tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/RPG/#p=0&tab=NewReleases",
                  "http://store.steampowered.com/tag/pt/Simula%C3%A7%C3%A3o/#p=0&tab=↩
                      NewReleases"
                  ]
    allowed_domains = ["store.steampowered.com"]
```

[1] Parsing de conteúdo HTML web para extração de informação.

Listing 2: Parser que irá ser aplicado ao conjunto de páginas alojadas nos URL mencionados anteriormente. Aplicação recursiva nas páginas associadas aos links presentes nestas.

```
def parse(self, response):
        for href in response.xpath("//a/@href[contains(.,'http://store.steampowered.com/app/')]"):
            app_url = re.compile("http://store.steampowered.com/app/(\d+)/")
            url = response.urljoin(href.extract())
            if app_url.match(url):
                app_id = app_url.search(url).group(1)
                if app_id not in self.visited_ids:
                    self.visited_ids.append(app_id)
                    more_like_url = 'http://store.steampowered.com/recommended/morelike/app/'+str(app_id)
                    yield scrapy.Request(url, callback=self.parse_dir_contents)
                    yield scrapy.Request(more_like_url, callback=self.parse)
                    yield scrapy.Request(url, callback=self.parse)
```

Listing 3: Exemplo de captura do atributo $nome$ de cada jogo recorrendo a *XPath* sobre a árvore *HTML*.

```
#NOME DO JOGO
        for res in response.xpath("//div[@class='apphub_AppName']").xpath("text()").extract():
            item['nome'] = res
```

Listing 4: Comando que exercita o código acima descrito e produz um documento *XML* resultante.

```
>> scrapy crawl steam_ontologia  o  resultado.xml
```

2.2 A ontologia

Baseado na informação disponível/obtida no processo acima descrito, estipulou-se que a ontologia seria composta por 6 classes com a disposição ilustrada na figura 1.

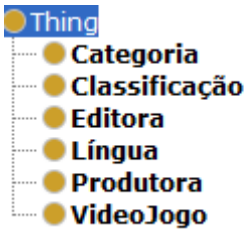

Figura 1: Disposição das classes na ontologia

Para além das relações inter classe, existem diversas *data properties* como exposto na figura 2. As relações entre classes são ilustradas na figura 3, como *object properties*.

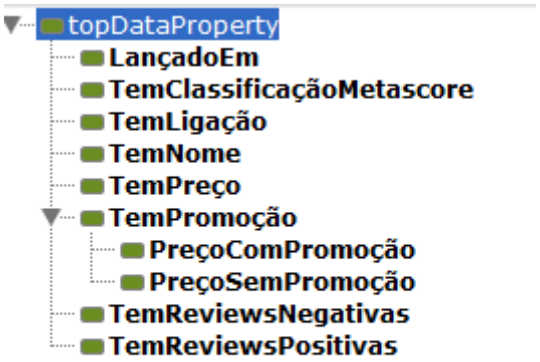

Figura 2: *Data properties*

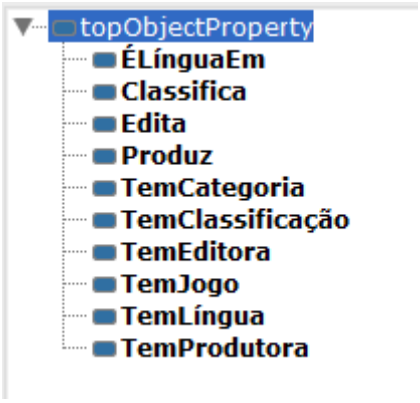

Figura 3: *Object properties*

2.3 De *XML* para *OWL*

Adquiridos os dados e definida a ontologia, foi então necessário povoá-la. Para isto, utilizou-se um documento *XSL* que reestruturou a informação extraída previamente e a converteu num documento *OWL* apto a ser executado na ferramenta *Protégé*.

2.4 Queries

Para confirmar o sucesso do porte de informação, assim como a utilidade do produto construído, produziu-se a seguinte query *SPARQL*:

Listing 5: Query SPARQL que identifica os cinco jogos mais baratos.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX so: <http://www.semanticweb.org/jpereira/ontologies/2015/6/steam-ontologia#>


SELECT ?nome ?preco
WHERE { ?jogo rdf:type so:VideoJogo.
    ?jogo so:TemNome ?nome.
    ?jogo so:TemPreco ?preco.
}
ORDER BY ?preco
LIMIT 5
```

A figura 4, é o resultado de uma captura de ecrã sobre a página HTML correspondente ao item mais barato. Como se pode verificar, os valores coincidem.

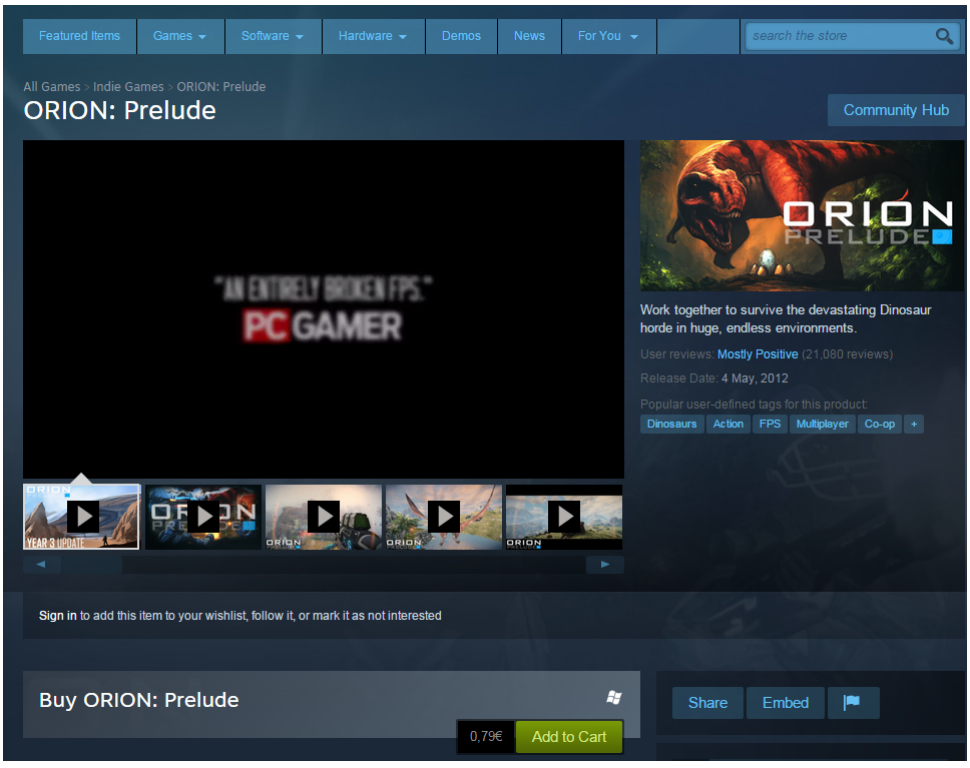

Figura 4: Página HTML do jogo mais barato

3 Conclusão

Ontologias são conceitos muito poderosos e aplicáveis a qualquer contexto e que possibilitam inferir novo conhecimento útil baseado em regras algébricas simples.
Para além disto, foi enriquecedor manipular as ferramentas descritas, marcando referências para trabalho futuro.