Construção automática de ontologias e sua utilização em sistemas de recuperação de informação em texto

José Saias and Paulo Quaresma

Departamento de Informática, Universidade de Évora,
7000 Évora, Portugal, jsaias|pq@di.uevora.pt

16 de Setembro de 2002

Sumário

De modo a serem mais poderosos, os sistemas de recuperação de informação em bases de texto necessitam de ter a capacidade de representar e de usar informação semântica sobre o conteúdo desses documentos.

Neste artigo é descrita a metodologia utilizada para transformar um sistema clássico de recuperação de informação em bases de textos num sistema de recuperação em bases de documentos com informação semântica. O processo de construção semi-automática da ontologia necessária à representação do conhecimento também é descrito.

1 Introdução

Em trabalho anterior Quaresma e Rodrigues [2] descreveram o sistema de recuperação de informação da Procuradoria Geral da República Portuguesa. Este sistema, disponível na web (http://www.pgr.pt), tem cerca de 7,000 documentos e cerca de 10,000,000 de palavras. Os documentos têm um estrutura específica e encontram-se definidos num formato compatível com XML.

O objectivo deste trabalho é descrever a metodologia aplicada para transformar o sistema de RI (Recuperação de Informação) existente num sistema com capacidade de representar conhecimento semântico e de o usar em processos de inferência.

A metodologia pode ser dividida nas seguintes fases:

- Definição de uma ontologia adequada;
- Transformação da base de textos em documentos com informação semântica;
- Desenvolvimento de um sistema de recuperação de informação com capacidade de processar e utilizar a ontologia definida e a informação semântica associada a cada documento.

É de salientar que a metodologia proposta permite satisfazer dois objectivos adicionais:

- Criar uma base de documentos compatível com a "semantic web", isto é, que possa ser acedida através de agentes existentes na web;
- Desenvolver um agente para recuperação de informação com capacidade de processar documentos numa linguagem "semantic web".

A secção 2 descreve o processo de criação da ontologia. A secção 3 descreve o processo de transformação dos documentos e a secção 4 descreve o sistema de recuperação de informação construído.

2 Criação da ontologia

O primeiro passo no processo de criação de uma ontologia adequada à base de textos existente é a escolha da linguagem de representação da ontologia.

Tendo em conta o requisito de que o sistema deverá poder ser disponibilizado através da web e que os documentos a representar se encontram em formato compatível com XML, a escolha óbvia recai nas linguagens chamadas "semantic web". De facto, nos últimos anos diversas linguagens definidas sobre o XML têm sido propostas para a representação de ontologias: RDF (Resource Description Framework - [1]), e DAML+OIL (Darpa Agent Markup Language - [3]) são algumas das mais relevantes. No âmbito deste trabalho optou-se pela linguagem DAML+OIL [3], dado possuir já um "standard" aprovado.

Com recurso a esta linguagem, havia que criar uma ontologia apropriada, isto é, uma especificação formal de como representar objectos, conceitos e entidades, com as respectivas características e relações que se estabelecem entre eles.

A estrutura da base de textos da Procuradoria Geral da República Portuguesa (PGR) foi analisada e foram identificadas duas classes de objectos fundamentais, com diversos "campos" ou atributos.

- Documento
- Conceito

A classe principal, *Documento*, possui um conjunto de campos bem delimitados, que vão corresponder a propriedades das suas instâncias. Efectivamente, cada documento original será representado através de uma instância da classe *Documento*.

Como exemplo, vejamos parte da definição daml+oil desta classe e das suas propriedades (neste caso, o facto dos documentos possuirem um determinado número identificador):

```
<daml:Class rdf:ID="Documento">
<daml:label>Documento</daml:label>
</daml:Class>

<daml:DatatypeProperty rdf:ID="numDoc">
<daml:domain rdf:resource="#Documento"/>
<rdf:type rdf:resource="http://www.daml.org
        /2001/03/daml+oil#UniqueProperty"/>
<daml:range rdf:resource="http://www.w3.org
        /2000/10/XMLSchema#string"/>
</daml:DatatypeProperty>
```

A outra classe importante é a classe *Conceito*, que serve para representar os conceitos, ou assuntos, que surgem no campo "descritores", dos documentos. Estes objectos têm propriedades que os relacionam com outros objectos do mesmo tipo: maisGeralQue, maisEspecíficoQue, conceitoRelacionado e conceitoEquivalente. Vejamos a definição daml+oil desta classe e de algumas das suas propriedade:

```
<daml:Class rdf:ID="Conceito">
<daml:label>Conceito</daml:label>
</daml:Class>

<daml:DatatypeProperty rdf:ID="nomeConceito">
<daml:domain rdf:resource="#Conceito"/>
<rdf:type rdf:resource="http://www.daml.org/
          2001/03/daml+oil#UniqueProperty"/>
<daml:range rdf:resource="http://www.w3.org/
         2000/10/XMLSchema#string"/>
</daml:DatatypeProperty>

<daml:ObjectProperty rdf:ID="maisGeralQue">
<daml:domain rdf:resource="#Conceito"/>
<daml:range rdf:resource="#Conceito"/>
</daml:ObjectProperty>

<daml:ObjectProperty rdf:ID="maisEspecificoQue">
<daml:inverseOf rdf:resource="#maisGeralQue"/>
</daml:ObjectProperty>
```

Como exemplo, vejamos a definição de um conceito (acidente):

```
<pgr:Conceito rdf:ID="c7276">
  <pgr:nomeConceito>Acidente</pgr:name>
  <pgr:maisGeralQue rdf:resource="#c1346"/>
  <pgr:maisEspecificoQue rdf:resource="#c7275"/>
</pgr:Conceito>
```

3 Transformação dos documentos

O passo seguinte na metodologia definida é transformar os documentos originais em documentos "semantic web" com formato daml+oil.

Para tal, foi construido um programa que automaticamente processa os dados dos ficheiros originais e que cria a representação daml+oil de acordo com a ontologia criada. Há dois tipos de ficheiros a processar:

- Conceitos
- Documentos

O primeiro passo foi construir um programa para representar os objectos da classe Conceito, com todas as relações para os outros conceitos. A hierarquia de conceitos encontrava-se previamente representada através de uma base de dados, tendo sido construído uma ferramenta que automaticamente exporta a base de dados para um ficheiro daml+oil.

De seguida, foi feito um programa para tratar cada documento, gerando a representação daml+oil respectiva, usando referências para os conceitos já criados.

A verificação do código gerado foi efectuada com uma ferramenta Daml+Oil (em Java) disponível em "http://www.daml.org/validator/".

4 Desenvolvimento de um sistema de recuperação de informação

O passo seguinte na metodologia proposta consiste no desenvolvimento de um sistema de recuperação de informação que permita efectuar interrogações sobre a representação semântica efectuada.

Como objectivo final, pretendia-se permitir interrogações do seguinte tipo:

- Quais os documentos onde a propriedade P=V? (V representa um determinado valor)
- Quais os documentos com o descritor D?
- Quais os documentos onde há um descritor do tipo T?

Um descritor do tipo T significa um conceito T ou um conceito C relacionado com T, onde T tem a propriedade "mais geral que" C, directa ou indirectamente.

Neste sentido, o motor de pesquisa a utilizar deveria ter capacidade para representar conhecimento e para efectuar inferências sobre esse conhecimento. Optou-se por utilizar a linguagem de programação declarativa Prolog.

Como consequência desta opção foi necessário construir um analisador daml+oil, que recebe como "input" um ficheiro daml+oil e que cria como "output" um conjunto de factos e regras Prolog. Dado que o ponto de partida deste trabalho é o código daml+oil, foi usado um parser XML para obter o valor dos campos.

Após a análise e conversão dos documentos daml+oil para factos e regras Prolog é possível efectuar interrogações a todos os campos das diversas instâncias como, por exemplo:

- Encontrar pareceres com uma determinada propriedade P igual a um dado valor V: parecerComP(P,V)
- Encontrar pareceres que tenham um dado descritor D: parecerComDescritor(P,D)
- Encontrar pareceres com um descritor do tipo T: parecerComDescritorTipo(P,T)

Em suma, através da aplicação da metodologia proposta, foi possível efectuar uma representação daml+oil dos documentos da PGR. Por outro lado, foi também elaborada uma representação em Prolog que permite utilizar o seu motor de inferência para efectuar interrogações.

Referências

[1] O. Lassila and R. Swick. *Resource Description Framework (RDF) - Model and Syntax Specification.* W3C, 1999.

[2] Paulo Quaresma and Irene Pimenta Rodrigues. PGR: Portuguese atorney general's office decisions on the web. In Osamu Yoshie, editor, *Proceedings of the 14th International Conference on Applications of Prolog*, University of Tokyo, Tokyo, Japan, October 2001. REN Associates, Inc. ISSN 1345-0980. To be published by Springer Verlag's LNAI.

[3] www.daml.org. *DAML+OIL – DARPA Agent Markup Language*, 2000.